An Algebraic Approach to Structural Learning Bayesian Networks

Milan Studený
Institute of Information Theory and Automation
Prague, Pod vodárenskou věží 4, CZ 18208, Czech Republic
e-mail: studeny@utia.cas.cz

Abstract

The basic idea of the paper is that every Bayesian network (BN) model is uniquely described by a certain integral vector, named a *standard imset*. Every score equivalent decomposable criterion $\mathcal{Q}$ for learning BN models appears to be an affine function of the standard imset. The algebraic view can naturally be extended to databases: if a criterion $\mathcal{Q}$ of the above mentioned kind is fixed then every database can be represented in the form of a *data vector* (relative to $\mathcal{Q}$), which is a vector of the same dimension as the standard imset.

Keywords: Standard Imset, Learning Bayesian nets, Quality Criterion.

1 INTRODUCTION

The procedures for learning *Bayesian network* (BN) models [7] can roughly be divided into two basic groups. Some of the algorithms are based on significance tests, that is, statistical *conditional independence* (CI) tests. The second basic group of algorithms consists of the procedures based on the maximization of a suitable *quality criterion*. The algorithms of this kind became popular in recent years. The present paper deals with structural learning based on the maximization of a quality criterion by the *local search method* – see [4].

The goal of the paper is to present briefly an algebraic approach to this special learning method. It brings surprisingly clear perspective on the method, which can possibly be extended to a structural learning method for general CI models. Actually, the presented algebraic approach is an attempt to apply a general algebraic method for describing probabilistic CI structures from [10] to learning BN models. The basic idea is that every BN model – that is, the class of distributions satisfying the CI restrictions determined by an acyclic directed graph – can be represented by a certain vector, whose components are integers, named a *standard imset*. This vector is uniquely determined representative of the BN model, like the well-known essential graph [1]. The standard imset corresponding to an acyclic directed graph G will be denoted by u_G.

The common criteria used in practice in the local search method typically satisfy two basic requirements: they are *score equivalent* [2] and *decomposable* [4]. These two requirements imply together that the criterion is necessarily a shifted linear function of the standard imset. More precisely, provided $\mathcal{Q}$ is a quality criterion of this kind, it has the form

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle \,, \qquad (1)$$

where G is a graph, D a database, $\langle x, y \rangle$ denotes the scalar product of vectors x and y, $s_D^{\mathcal{Q}}$ is a real constant and $t_D^{\mathcal{Q}}$ a vector depending on the database, called the *data vector* (relative to $\mathcal{Q}$). Thus, all substantial information about the database is involved in the data vector and the problem of maximization of $\mathcal{Q}$ is transformed to the problem of maximization of a linear function (determined by

t_D^Q) on a finite set of vectors, namely on the set of standard imsets.

A further remarkable property of standard imsets is as follows. In the local search method, it appears to be advantageous to consider the neighborhood structure for BN models derived from the inclusion of BN models – which is named *inclusion* (boundary) *neighborhood* [6]. It is a pleasant property of the algebraic approach that the neighborhood of two BN models in this sense can be characterized in terms of the respective standard imsets. More specifically, it is equivalent to the condition that their difference, named the *differential imset* is a certain simple vector that corresponds to an elementary CI statement. These vectors are named *elementary imsets* for their exceptional role in the method reported in [10].

Standard imsets are also appropriate to testing of the inclusion of BN models. Note that a graphical transformational characterization of inclusion of BN models was given by [4]. However, the algebraic characterization seems to be more elegant. Two BN models are in inclusion iff their differential imset is a combination of elementary imsets, named a *combinatorial imset*. Note that the question of computer testing whether a given vector is a combinatorial imset is very close to the problem of computer testing CI inference dealt with in [3]. Altogether, the algebraic approach leads to a proposal to modify the local search method so that some linear programming algorithms could be utilized in the future.

The structure of the paper is as follows. In Section 2 some basic concepts are introduced. Section 3 is a review of the local search method for the maximization of a quality criterion. We recall the concept of a score equivalent decomposable criterion (for learning BN models) and the concept of inclusion neighborhood. The algebraic approach is presented in Section 4. The concept of a standard imset is introduced and the above mentioned claims are formulated in a precise way. The idea of modification of the local search method in light of the algebraic approach is pinpointed in the Conclusion. Some open problems are mentioned there as well. Note that the proofs of the results mentioned in the present paper can be found in Chapters 7 and 8 of [10].

2 BASIC CONCEPTS

Throughout the paper the symbol N will denote a non-empty finite set of variables. They correspond to primitive factors described by random variables. Every BN model over N is given by an acyclic directed graph having N as the set of nodes. The class of all these graphs will be denoted by $\mathsf{DAGS}(N)$. Given $G \in \mathsf{DAGS}(N)$ and $i \in N$ the symbol $pa_G(i)$ will denote the set of *parents* of the node i, that is, the set of $j \in N$ with $j \to i$ in G.

2.1 PARAMETRIZATION OF A BAYESIAN NETWORK MODEL

A BN model is understood as a statistical model, that is, a class of probability distributions. To specify it the respective sample spaces have to be fixed. In this paper, the discrete case is only considered: let X_i denote a non-empty finite set of possible values for a variable $i \in N$. Given $A \subseteq N$, the symbol X_A will denote the set of *configurations* of values for A, that is, the set of lists $[x_i]_{i \in A}$ where $x_i \in \mathsf{X}_i$ for any $i \in A$.[1] Given $x \in \mathsf{X}_N$ and $A \subseteq N$, the symbol x_A will denote the *marginal configuration* of x for A, that is, the restriction of the list x to items indexed by elements of A.

A probability distribution P on X_N is given by its *density* p, which is a function $p : \mathsf{X}_N \to [0,1]$ with $\sum \{p(x);\, x \in \mathsf{X}_N\} = 1$. The *BN model* given by $G \in \mathsf{DAGS}(N)$ consists of the class of *Markovian distributions* with respect to G, that is, those distributions on X_N which satisfy CI restrictions dictated by the d-separation criterion – for detail see [8]. It will be denoted by $\mathbb{M}_G$. One can interpret $\mathbb{M}_G$ as a parameterized class of distributions. To describe the "standard" parameterization of $\mathbb{M}_G$ some conventions are needed.

[1] Of course, if $A \neq \emptyset$ then X_A is the Cartesian product $\prod_{i \in A} \mathsf{X}_i$. The set $\mathsf{X}_\emptyset$ consists of one configuration, namely the empty list.

CONVENTION 1 *The letter i will be used as a generic symbol for a variable: $i \in N$. We put $r(i) = |\mathsf{X}_i|$ and fix an ordering $y_i^1, \ldots, y_i^{r(i)}$ of elements of X_i for every $i \in N$. The letter k will be used as a generic symbol for a code of a configuration in X_i: $k \in \{1, \ldots, r(i)\}$. Given $i \in A \subseteq N$ and $x \in \mathsf{X}_A$ the symbol $k(i,x)$ will denote the code of $x_{\{i\}}$, that is, the unique $1 \leq k \leq r(i)$ such that $x_{\{i\}} = y_i^k$.*

Given $G \in \mathsf{DAGS}(N)$, denote by $q(i,G)$ the number $|\mathsf{X}_{pa_G(i)}|$ of parent configurations for a variable $i \in N$. Of course, $q(i,G) = \prod_{\ell \in pa_G(i)} r(\ell)$. Fix an ordering $z_i^1, \ldots, z_i^{q(i,G)}$ of elements of $\mathsf{X}_{pa_G(i)}$ for every $i \in N$. The letter j will be used as a generic symbol for a code of a parent configuration: $j \in \{1, \ldots, q(i,G)\}$. Given $i \in N$ such that $pa_G(i) \subseteq A \subseteq N$ and $x \in \mathsf{X}_A$ the symbol $j(i,x)$ will denote the code of $x_{pa_G(i)}$, that is, the unique $1 \leq j \leq q(i,G)$ such that $x_{pa_G(i)} = z_i^j$.

The set of parameters Θ_G for $\mathbb{M}_G$ consists of vectors $\boldsymbol{\theta} = [\theta_{ijk}]$ where $\theta_{ijk} \in [0,1]$ for $i \in N$, $1 \leq j \leq q(i,G)$ and $1 \leq k \leq r(i)$ such that $\sum_{k=1}^{r(i)} \theta_{ijk} = 1$ for every $i \in N$ and $1 \leq j \leq q(i,G)$. Given a vector parameter $\boldsymbol{\theta}$ the respective distribution is given by its density

$$p^{\boldsymbol{\theta}}(x) = \prod_{i \in N} \theta_{i\, j(i,x)\, k(i,x)} \quad \text{for } x \in \mathsf{X}_N\,. \quad (2)$$

It is shown in Lemma 8.1 of [10] that $p^{\boldsymbol{\theta}}$ is always a density of a distribution $P^{\boldsymbol{\theta}}$ on X_N and the mapping $\boldsymbol{\theta} \mapsto P^{\boldsymbol{\theta}}$ is onto $\mathbb{M}_G$.

2.2 EQUIVALENCE OF GRAPHS

Two acyclic directed graph $G, H \in \mathsf{DAGS}(N)$ are called *Markov equivalent* if $\mathbb{M}_G = \mathbb{M}_H$. If we consider non-degenerate sample spaces, that is, $|\mathsf{X}_i| \geq 2$ for every $i \in N$, then this is equivalent to the condition that G and H are *independence equivalent*, by which is meant that they define the same collection of CI restrictions through the d-separation criterion.[2] We write $G \approx H$ then. A well-known graphical characterization of independence equivalence is that $G \approx H$ iff G and H have the same underlying undirected graph and immoralities[3] – for a proof see [1].

[2] One can use Theorem 8.3 in [7].

2.3 DATA

By a *database* over N of the length d, $d \in \mathbb{N}$ will be understood a sequence $x^1, \ldots, x^d$ of elements of X_N.[4] Thus, complete databases are only considered here. The class of databases over N of the length d will be denoted by $\mathsf{DATA}(N,d)$. To give elegant formulas for basic quality criteria an additional convention is needed.

CONVENTION 2 *Keeping Convention 1 in mind let $D : x^1, \ldots, x^d$, $d \geq 1$ be a database. Introduce for every $i \in N$, $1 \leq j \leq q(i,G)$:*

$$d_{ij} = |\{1 \leq \ell \leq d;\, x^\ell_{pa_G(i)} = z_i^j\}|\,.$$

If, moreover, $1 \leq k \leq r(i)$ then put

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

Note that the order of configurations in a database is often considered not to be important.

3 LOCAL SEARCH METHOD

In this section, the idea of learning BN models based on the maximization of a quality criterion is described.

3.1 QUALITY CRITERION

By a *quality criterion* (for learning BN models) is meant a function $\mathcal{Q} : \mathsf{DAGS}(N) \times \mathsf{DATA}(N,d) \to \mathbb{R}$, $d \geq 1$. The reader can find several alternative phrases in the literature, e.g. quality measure [2] and scoring criterion [4, 7]. The intention of a learning procedure is, given a database $D \in \mathsf{DATA}(N,d)$, to find $G \in \mathsf{DAGS}(N)$ which maximizes the function $G \mapsto \mathcal{Q}(G,D)$. Of course, what is

[3] An *immorality* is an induced subgraph $a \to c \leftarrow b$, that is, $[a,b]$ is not an edge in the graph.

[4] Note that I intentionally introduce the concept of a database as an empirical concept, that is, a concept introduced in terms of observed evidence only. It does not involve statistical assumptions on data generating process. In my view, these additional details only complicate a clear view on the subject.

written here is a pure mathematical concept and there are a number of additional implicit assumptions on a quality criterion.

One of them is the assumption of *consistency* – see § 8.4.2 in [7]. The intuitive meaning of this concept is as follows: if a database is "generated" by a distribution P over N then the maximum of $\mathcal{Q}$ is achieved in the "simplest" $G \in \mathsf{DAGS}(N)$ with $P \in \mathbb{M}_G$. This is a natural requirement and most of the criteria used in practice satisfy this basic statistical assumption.

Let us give some examples of quality criteria. A classic statistical interpretation of graphical models as parameterized classes of probability distributions leads to *information criteria*. These are derived from maximized likelihood – see § 11.3.1 in [5]. Having fixed a BN model $\mathbb{M}_G$, $G \in \mathsf{DAGS}(N)$ the *likelihood function* ascribes to every database $D \in \mathsf{DATA}(N, d)$ and to a parameter $\boldsymbol{\theta} \in \Theta_G$ the probability of occurrence of D provided that data are "generated" from $P^{\boldsymbol{\theta}}$. The model $\mathbb{M}_G$ can then be "evaluated" by the maximum of the logarithm of the likelihood function. Provided that one has in mind the parameterization from § 2.1, the respective *maximized log-likelihood criterion* has the following form – see Corollary 8.1 in [10]:

$$\mathsf{MLL}(G, D) \;=\; \sum_{i \in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \frac{d_{ijk}}{d_{ij}}\,, \quad (3)$$

where $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$ and the convention $0 \cdot \ln(0/\star) \equiv 0$ is accepted.

However, this criterion does not take into account the simplicity (or complexity) of a BN model. The complexity of $\mathbb{M}_G$ can be measured by its *effective dimension*, that is, the number of free parameters in Θ_G:

$$\mathsf{DIM}\,(G) = \sum_{i \in N} q(i, G) \cdot [r(i) - 1]\,. \quad (4)$$

Information criteria can be obtained from the maximized log-likelihood criterion by subtracting a penalization term, which is typically a multiple of the effective dimension. The simplest one is *Akaike's information criterion*:

$$\mathsf{AIC}(G, D) \;=\; \mathsf{MLL}(G, D) - \mathsf{DIM}(G).$$

The most popular is probably *Bayesian information criterion*:

$$\mathsf{BIC}(G, D) \;=\; \mathsf{MLL}(G, D) - \frac{\ln d}{2} \cdot \mathsf{DIM}(G)\,.$$

This criterion is claimed to be consistent – see § 8.4.3 in [7]. Another class of allegedly consistent criteria is the class of various *Bayesian criteria*.[5] These are derived from marginal likelihood – see § 11.3.3 in [5]. The basic idea is that a prior probability distribution π_G on the parameter space Θ_G is considered for every $G \in \mathsf{DAGS}(N)$. The respective Bayesian criterion is the *logarithm of the marginal likelihood*:

$$\mathsf{LML}(G, D) \;=\; \ln \int_{\Theta_G} L(\boldsymbol{\theta}, D)\, \mathrm{d}\, \pi_G(\boldsymbol{\theta})\,,$$

where $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$ and $L(\boldsymbol{\theta}, D)$ denotes the value of the likelihood function for $\boldsymbol{\theta}$ and D.

3.1.1 Score equivalent criteria

A quality criterion $\mathcal{Q}$ will be called *score equivalent* if, for every $G, H \in \mathsf{DAGS}(N)$ and each $D \in \mathsf{DATA}(N, d)$,

$$\mathcal{Q}(G, D) \;=\; \mathcal{Q}(H, D) \quad \text{whenever } G \approx H.$$

This basically means that whenever G and H define the same BN model then the criterion gives them the same "score", no matter what database is considered. This is a natural requirement from a statistical point of view.

Note that quality criteria used in practice are usually score equivalent. These are both Akaike's and Bayesian information criteria – see Proposition 8.2 in [10]. Most of Bayesian criteria are also score equivalent, but there is an example of a Bayesian criterion which is not score equivalent, namely so-called K2 metric – see [2, 6].

[5] The reviewer objects my using the word "allegedly". However, I myself have not found the arguments given in § 8.4.3 of [7] convincing enough; they are too sketchy and based on references. Although I tend to believe what is claimed I was not able to really check it (from the perspective of a mathematician).

3.1.2 Decomposable criteria

To introduce the concept of a decomposable criterion one needs to know what is a projection of a database. Given a database $D : x^1, \ldots, x^d$, $d \geq 1$ over N and $A \subseteq N$, its *projection* D_A onto A is the sequence of the respective marginal configurations $x^1_A, \ldots, x^d_A$.

A quality criterion is *decomposable* if there exist functions $q_{i|B} : \mathsf{DATA}(\{i\} \cup B, D) \to \mathbb{R}$, where $i \in N$, $B \subseteq N \setminus \{i\}$, $d \geq 1$ such that

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

for $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$. Informally said, a criterion is decomposable if it decomposes recursively with respect to the graph. The assumption of decomposability was introduced with connection to the local search method. It is a technical assumption which allows one to search for a local maximum of $\mathcal{Q}$ by this method.

3.2 THE IDEA OF LOCAL SEARCH

The problem with the maximization of a quality criterion is that the set $\mathsf{DAGS}(N)$, respectively the collection of equivalence classes $\mathsf{DAGS}(N)/_\approx$, is too large and this makes direct maximization infeasible. To circumvent this task the method of local search was proposed. The basic idea is to introduce in the *search space*, which is either the set of graphs $\mathsf{DAGS}(N)$ or the set $\mathsf{DAGS}(N)/_\approx$, a neighborhood structure. Every graph, respectively every equivalence class, is assigned a relatively small set of neighbors; they typically differ in the presence of one edge. Instead of looking for a global maximizer of $\mathcal{Q}$ one searches for its local maximizer relatively to the chosen neighborhood structure. The point is that, for a decomposable criterion $\mathcal{Q}$, the difference in the value of $\mathcal{Q}$ for neighbors is often easy to compute. Thus, the method consists in travelling in the search space. In each *state*, that is, in each element of the search space, one considers a limited collection of possible *moves* to neighboring states. Each time one chooses the move that maximizes the increase in the value of $\mathcal{Q}$.

3.3 INCLUSION NEIGHBORHOOD

A natural neighborhood structure for the search space $\mathsf{DAGS}(N)/_\approx$ is the one derived from the inclusion of BN models. Let $\mathcal{M}_G$ denote the collection of CI restrictions determined by $G \in \mathsf{DAGS}(N)$ through the d-separation criterion. Given $K, L \in \mathsf{DAGS}(N)$ we say that K is *independence included* in L if $\mathcal{M}_K \subseteq \mathcal{M}_L$. Note that, if we consider non-degenerate sample spaces, this is equivalent to $\mathbb{M}_L \subseteq \mathbb{M}_K$, that is, L is *distributionally included* in K – see § 8.4.1 in [7].

The symbol $\mathcal{M}_K \subset \mathcal{M}_L$ denotes strict inclusion, that is, $\mathcal{M}_K \subseteq \mathcal{M}_L$ but $\mathcal{M}_K \neq \mathcal{M}_L$. If $\mathcal{M}_K \subset \mathcal{M}_L$ and, moreover, there is no $G \in \mathsf{DAGS}(N)$ with $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$ then we say that $\mathcal{M}_L$ is an *upper inclusion neighbor* of $\mathcal{M}_K$, respectively $\mathcal{M}_K$ is a *lower inclusion neighbor* of $\mathcal{M}_L$. Then we will write $\mathcal{M}_K \sqsubset \mathcal{M}_L$. Clearly, the inclusion neighborhood has good theoretical justification. Nevertheless, there are also some practical reasons why reasonable neighborhood structure for $\mathsf{DAGS}(N)/_\approx$ should involve the inclusion neighborhood – see [6].

A graphical characterization of inclusion neighborhood relation is as follows (see Lemma 8.5 in [10]): $\mathcal{M}_K \sqsubset \mathcal{M}_L$ occurs iff there exist $K', L' \in \mathsf{DAGS}(N)$ with $K' \approx K$, $L' \approx L$ such that L' is obtained from K' by an arrow removal.

3.4 PROBLEM OF REPRESENTATIVE CHOICE

One of the issues related to the implementation of the local search method is the problem of representing a BN model in the memory of a computer. Some researchers prefer to represent a BN model by any graph in the respective equivalence class $\mathsf{DAGS}(N)/_\approx$. This, however, may lead to computational inefficiencies since these equivalence classes could be quite large and the procedure can stick at idle graphical operations. Other researchers prefer to use unique graphical representatives. Since, in general, there is no distinguished member of an equivalence class of acyclic directed graphs, the authors use special chain

graphs for this purpose. The most popular representative of this kind is the *essential graph* [1], named also the completed p-dag in [4]. The basic idea of this paper is to represent a BN model by an algebraic representative.

4 ALGEBRAIC VIEW

The above mentioned algebraic representatives of BN models will be special vectors. Let $\mathcal{P}(N)$ denote the power set of the set of variables, that is, $\mathcal{P}(N) = \{A;\, A \subseteq N\}$. By an *imset* over N an integer-valued function on $\mathcal{P}(N)$ will be understood.[6] Of course, an imset over N can be viewed as an element of $\mathbb{Z}^{\mathcal{P}(N)}$, that is, an integral vector, whose components are indexed by subsets of N. Arithmetic operations with imsets are defined coordinate-wise. This allows one to express every imset as a linear combination of elements of a linear base of $\mathbb{R}^{\mathcal{P}(N)}$. The usual base of $\mathbb{R}^{\mathcal{P}(N)}$ consists of vectors that identify subsets of N. Given $A \subseteq N$, the symbol δ_A will denote the following imset:

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

In [10] a special class of imsets was proposed to describe probabilistic CI structures. These imsets can be introduced as certain combinations of imsets that correspond to elementary CI statements. Thus, every triplet $\langle a, b|C\rangle$, where $a, b \in N$ are distinct and $C \subseteq N \setminus \{a, b\}$, defines the respective *elementary imset* $u_{\langle a,b|C\rangle}$[7] by the formula:

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

A *combinatorial imset* is an imset which can be obtained as a sum of elementary imsets while their repetition is allowed. Every discrete CI structure can perfectly be described by a combinatorial imset – see Theorem 5.2 and Corollary 5.3 in [10].[8]

[6] The word "imset" is an abbreviation for **i**nteger-valued **m**ulti**set**.

[7] This imset corresponds to the CI statement "a is conditionally independent of b given C", in notation $a \perp\!\!\!\perp b \,|\, C$.

[8] Note for explanation that there exists an algebraic criterion determining CI restrictions dictated by an imset of this kind – see § 4.4.1 in [10]. This criterion can be viewed as an analog of the graphical d-separation criterion from [8].

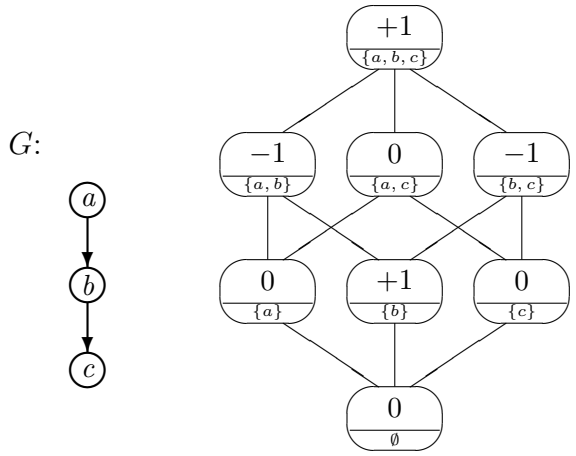

Figure 1: A graph and a standard imset.

4.1 STANDARD IMSETS

Given $G \in \mathsf{DAGS}(N)$, the respective *standard imset* u_G is given by the following formula:

$$u_G \;=\; \delta_N - \delta_\emptyset + \sum_{i \in N} \{\, \delta_{pa_G(i)} - \delta_{\{i\}\cup pa_G(i)} \,\}. \quad (5)$$

It is shown in Lemma 7.1 in [10] that the standard imset u_G is always a combinatorial imset and defines the same collection of CI restrictions as the graph G through the d-separation criterion. To illustrate this concept consider the graph G over $N = \{a, b, c\}$ shown on the left-hand side of Figure 1. The respective standard imset is shown in the diagram on the right-hand side of the figure.

Note that it follows from the definition that every standard imset over N has at most $2|N| + 2$ non-zero components. This can be utilized for effective representation standard imsets in the memory of a computer.

4.1.1 Equivalence characterization

The basic observation is as follows – for a proof see Corollary 7.1 in [10].

PROPOSITION 4.1 *Given $K, L \in \mathsf{DAGS}(N)$ one has $K \approx L$ iff $u_K = u_L$.*

In particular, the standard imset is a unique representative of the respective equivalence class of acyclic directed graphs, and, therefore, of the respective BN model. Note that

there exists a direct formula for the standard imset on basis the essential graph and an inverse reconstruction algorithm – see [9].

4.1.2 Inclusion characterization

Another advantage of standard imsets is as follows.

PROPOSITION 4.2 *Assume $K, L \in \mathsf{DAGS}(N)$. Then $\mathcal{M}_K \subseteq \mathcal{M}_L$ iff $u_L - u_K$ is a combinatorial imset. Moreover, one has $\mathcal{M}_K \sqsubset \mathcal{M}_L$ iff $u_L - u_K$ is an elementary imset.*

The proof can be found in § 8.4.1 of [10]. The characterization of inclusion neighborhood relationship is straightforward since the recognition elementary imsets is immediate. Supposing $\mathcal{M}_K \sqsubset \mathcal{M}_L$ the unique elementary imset $u_L - u_K$ will be called the *differential imset* for K and L.

4.2 FORMULAS FOR CRITERIA

The main result of the paper is as follows.

THEOREM 4.1 *Let $\mathcal{Q}$ be a quality criterion for learning BN models which is both score equivalent and decomposable. Then every database $D \in \mathsf{DATA}(N, d)$, $d \geq 1$ can be assigned a number $s_D^{\mathcal{Q}} \in \mathbb{R}$ and a function $t_D^{\mathcal{Q}} : \mathcal{P}(N) \to \mathbb{R}$ such that, for every $A \subseteq N$, $t_D^{\mathcal{Q}}(A)$ only depends on the projection D_A of the database and*

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \sum_{A \subseteq N} t_D^{\mathcal{Q}}(A) \cdot u_G(A) \qquad (6)$$

for $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N, d)$. [9]

For the proof see Lemmas 8.7 and 8.3 in [10]. Of course, the function $t_D^{\mathcal{Q}}$ can be interpreted as a real vector $[t_D^{\mathcal{Q}}(A)]_{A \subseteq N} \in \mathbb{R}^{\mathcal{P}(N)}$, called the *data vector* (for the database D relative to $\mathcal{Q}$). Then the sum in (6) is the scalar product of the data vector and the standard imset. Thus, (6) is nothing but the formula (1) mentioned in the Introduction.

Let us illustrate the result by examples. The maximized log-likelihood criterion given by (3) is an example of a decomposable criterion. Let $\hat{P}$ denote the empirical distribution computed from D with density

$$\hat{p}(y) = d^{-1} \cdot |\{1 \leq \ell \leq d;\, x^\ell = y\}| \quad \text{for } y \in \mathsf{X}_N \,.$$

Then s_D^{MLL} is the $(-d)$-multiple of the entropy of $\hat{P}$:

$$s_D^{\mathsf{MLL}} \,=\, d \cdot \sum_{y \in \mathsf{X}_N, \hat{p}(y) > 0} \hat{p}(y) \cdot \ln \hat{p}(y) \,.$$

Given $A \subseteq N$, the value of $t_D^{\mathsf{MLL}}(A)$ is the d-multiple of the multiinformation of the marginal $\hat{P}_A$ of $\hat{P}$ on X_A:[10]

$$t_D^{\mathsf{MLL}}(A) \,=\, d \cdot H(\hat{P}_A | \prod_{i \in A} \hat{P}_i) \quad \text{for } A \subseteq N \,.$$

For a proof see Proposition 8.4 in [10]. The effective dimension (4) can be viewed as a special criterion which does not depend on D. Thus, Corollary 8.6 in [10] gives $s^{\mathsf{DIM}} = -1 + \prod_{i \in N} r(i)$ and

$$t^{\mathsf{DIM}}(A) \,=\, |A| - 1 + \prod_{i \in A} r(i) - \sum_{i \in A} r(i)$$

for $A \subseteq N$. Hence, the respective data vector for the Bayesian information criterion is

$$t_D^{\mathsf{BIC}}(A) \,=\, t_D^{\mathsf{MLL}}(A) - \frac{\ln d}{2} \cdot t^{\mathsf{DIM}}(A) \quad \text{for } A \subseteq N \,.$$

Theorem 4.1 has the following easy consequence which explains the role of the differential imset.

COROLLARY 4.1 *Given $K, L \in \mathsf{DAGS}(N)$ such that $\mathcal{M}_K \sqsubset \mathcal{M}_L$, let $u_{\langle a,b|C \rangle}$ be the differential imset for K and L. Then*

$$\mathcal{Q}(K, D) - \mathcal{Q}(L, D) = \langle t_D^{\mathcal{Q}}, u_{\langle a,b|C \rangle} \rangle$$

for every $D \in \mathsf{DATA}(N, d)$ and a score equivalent decomposable quality criterion $\mathcal{Q}$.

[9] Provided one accepts a standardization convention $t_D^{\mathcal{Q}}(A) = 0$ whenever $|A| \leq 1$, the numbers $s_D^{\mathcal{Q}}$ and $t_D^{\mathcal{Q}}(A)$, $A \subseteq N$ are uniquely determined.

[10] Recall that the multiinformation of a distribution Q is the relative entropy of Q with respect to the product $R = \prod_i Q_i$ of its one-dimensional marginals: $H(Q|R) = \sum_{y, q(y) > 0} q(y) \cdot \ln [q(y)/r(y)]$ where q, r are densities of Q, R.

CONCLUSIONS

The algebraic approach presented in this paper leads to the following proposal of how to modify the method of local search described in § 3.2. The states of the search space could be *standard imsets* and the moves between states can be represented by *differential imsets*. Given a quality criterion $\mathcal{Q}$ every database can be represented by the respective *data vector*.

Thus, graphical representatives of BN models are replaced by algebraic ones and, moreover, an algebraic representative of a database is incorporated. The algebraic interpretation of moves is possible owing to Corollary 4.1; note that an important fact is that every move of this kind has CI interpretation.

Of course, there are several open problems related to this topic. One of them is what are mutual geometric positions of standard imsets in $\mathbb{R}^{\mathcal{P}(N)}$. It is desirable to confirm the hypothesis that every standard imset is an extreme point of the polytope consisting of convex combinations of standard imsets. If this is true then the methods of linear programming can possibly be applied in this area.

Another open problem is to characterize, in algebraic terms, all the moves from a given standard imset to its inclusion neighbors. Note that these moves were already characterized in terms of the essential graph.

Acknowledgements

This research has been supported by the grant GAČR n. 201/04/0393.

References

[1] 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.

[2] R.R. Bouckaert (1995). Bayesian belief networks: from construction to evidence. PhD thesis, University of Utrecht.

[3] 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.

[4] D.M. Chickering (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research* **3**: 507-554.

[5] R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems.* New York: Springer-Verlag

[6] T. Kočka and R. Castelo (2001). Improved learning Bayesian networks. In *Uncertainty in Artificial Intelligence 17* (J. Breese, D. Koller eds.), Morgan Kaufmann, 269-276.

[7] R.E. Neapolitan (2004). *Learning Bayesian Networks.* New York: Pearson Prentice Hall.

[8] J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo: Morgan Kaufmann.

[9] 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.

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

[11] M. Studený (2005). Characterization of inclusion neighbourhood in terms of the essential graph. *International Journal of Approximate Reasoning* **38**: 283-309.