Characteristic imset: a simple algebraic representative of a Bayesian network structure

Milan Studený
Institute of Information Theory and Automation of the ASCR, Czech Republic
studeny@utia.cas.cz

Raymond Hemmecke
TU Munich, Germany
hemmecke@ma.tum.de

Silvia Lindner
University of Magdeburg, Germany
lindner@mail.math.uni-magdeburg.de

Abstract

First, we recall the basic idea of an algebraic and geometric approach to learning a Bayesian network (BN) structure proposed in (Studený, Vomlel and Hemmecke, 2010): to represent every BN structure by a certain uniquely determined vector. The original proposal was to use a so-called *standard imset* which is a vector having integers as components, as an algebraic representative of a BN structure. In this paper we propose an even simpler algebraic representative called the *characteristic imset*. It is 0-1-vector obtained from the standard imset by an affine transformation. This implies that every reasonable quality criterion is an affine function of the characteristic imset. The characteristic imset is much closer to the graphical description: we establish a simple relation to any chain graph without flags that defines the BN structure. In particular, we are interested in the relation to the *essential graph*, which is a classic graphical BN structure representative. In the end, we discuss two special cases in which the use of characteristic imsets particularly simplifies things: learning decomposable models and (undirected) forests.

1 Introduction

The score and search method for learning Bayesian network (BN) structure from data consists in maximizing a *quality criterion* $\mathcal{Q}$, also named a *scoring criterion* or simply a *score* by other authors. It is a real function of the (acyclic directed) graph G and the observed database D. The value $\mathcal{Q}(G, D)$ measures how well the BN structure defined by G fits the database D.

Two important technical requirements on the criterion $\mathcal{Q}$ emerged in the literature in connection with computational methods dealing with this maximization task: $\mathcal{Q}$ should be *score equivalent* (Bouckaert, 1995) and (additively) *decomposable* (Chickering, 2002).

Another important question is how to represent the BN structure in the memory of a computer. It could be the case that different acyclic directed graphs are Markov *equivalent*, i.e., they define the same BN structure. A classic graphical characterization of equivalent graphs (Verma and Pearl, 1991) states that they are equivalent iff they have the same *adjacencies* and *immoralities*, which are special induced subgraphs. Representing a BN structure by any of the acyclic directed graphs defining it leads to a non-unique description causing later identification problems. Thus, researchers calling for methodological simplification proposed to use a unique representative for each individ-

Pp. 257–265 in *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010)*, edited by Petri Myllymäki, Teemu Roos and Tommi Jaakkola. HIIT Publications 2010–2.

ual BN structure. The classic unique graphical representative is the *essential graph* (Andersson, Madigan and Perlman, 1997).

The idea of an algebraic approach, introduced in Section § 8.4 of (Studený, 2005), is to use an algebraic representative, called the *standard imset.* It is a vector whose components are integers indexed by subsets of the set of variables (= nodes) N. Moreover, it is a unique BN structure representative and the memory demands for its computer representation are polynomial in $|N|$. The most important point, however, is: Every score equivalent and decomposable criterion $\mathcal{Q}$ is an affine function (= linear function plus a constant) of the standard imset. Specifically, given an acyclic directed graph G (over N) and a database D, we have

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

where $s_D^{\mathcal{Q}}$ is a constant depending on the database and where $\langle t_D^{\mathcal{Q}}, u_G \rangle$ is the scalar product of a vector depending on the database, called the *data vector* (relative to $\mathcal{Q}$), and of the standard imset u_G (for G). Note that there is a polynomial-time algorithm (in $|N|$) for the reconstruction of the essential graph from the standard imset (Studený and Vomlel, 2009).

The geometric view was introduced in the paper (Studený, Vomlel and Hemmecke, 2010), where it was shown that the set of standard imset (over a fixed set of variables N) is the set of vertices (= extreme points) of a certain polytope. In particular, the maximization of $\mathcal{Q}$ over acyclic directed graphs can be re-formulated as a classic *linear programming problem*, that is, optimizing a linear function over a polyhedron.[1]

In this paper, we propose an alternative algebraic representative of a BN structure, called the *characteristic imset.* It is a vector obtained from the standard imset by a one-to-one affine transformation that maps lattice points to lattice points (in both directions). Thus, every score equivalent and decomposable criterion is an affine function of the characteristic imset and the set of characteristic imsets is the set of vertices of a polytope. The characteristic imset has only zeros and ones as its components. Moreover, it is very close to the graphical description: some of its components with value one correspond to adjacencies. Immoralities can also be recognized in the graph(s) on the basis of the characteristic imset. We establish a simple relation of the characteristic imset to any chain graph (without flags) defining the BN structure. In particular, this makes it possible to get immediately the characteristic imset on the basis of the essential graph. We also consider the converse task of reconstructing the essential graph from the characteristic imset.

If we restrict ourselves to decomposable models (= BN structures defined by acyclic directed graphs without immoralities), then the characteristic imset has a quite simple form. The situation is particularly transparent in the case of (undirected) forests: the edges of the forest are in one-to-one correspondence with 1's in the characteristic imset. The polytope spanned by these special characteristic imsets has already been studied in matroid theory (Schrijver, 2003). Consequently, we can easily learn (undirected) tree structures, which give an elegant geometric interpretation to a classic heuristic procedure proposed by Chow and Liu (1968).

The structure of this paper is as follows. In Section 2 we recall some of the definitions and relevant results. In Section 3 we introduce the characteristic imset and derive the above mentioned observations on it. Section 4 is devoted to the reconstruction of the essential graph from the characteristic imset. Section 5 briefly outlines our results about learning undirected forests from (Hemmecke et al., 2010). In Conclusions we discuss further perspectives.

2 Basic concepts

2.1 Graphical concepts

Graphs considered in this paper have a finite non-empty set of nodes N and two types of edges: directed edges, called *arrows*, denoted like $i \to j$ or $j \leftarrow i$, and undirected edges. No multiple edges are allowed between two nodes. If there is an edge between nodes i and j, we say they are *adjacent.*

[1] Note that a polytope is simply a bounded polyhedron.

Given a graph G over N and a non-empty set of nodes $A \subseteq N$, the *induced subgraph* of G for A has just those edges in G having both end-nodes in A. An *immorality* in G is an induced subgraph (of G) for three nodes $\{a, b, c\}$ in which $a \to c \leftarrow b$ and a and b are not adjacent. A *flag* is another induced subgraph for $\{a, b, c\}$ in which $a \to b$, b and c are adjacent by an undirected edge and a and c are not adjacent.

A set of nodes $K \subseteq N$ is *complete* in G if every pair of distinct nodes in K is adjacent by an undirected edge. A maximal complete set is called a *clique*. A set $C \subseteq N$ is *connected* if every pair of distinct nodes in C is connected via an undirected path. Maximally connected sets are called *components*.

A graph is *directed* if it has no undirected edges. A directed graph G over N is called *acyclic* if it has no directed cycle, that is, a sequence of nodes $a_1, \ldots, a_n$, $n \geq 3$ with $a_i \to a_{i+1}$ for $i = 1, \ldots n$, under the convention $a_{n+1} \equiv a_1$. An equivalent definition is the existence of an ordering $b_1, \ldots, b_m$, $m \geq 1$, of all nodes in N which is consistent with the direction of arrows, that is, $b_i \to b_j$ in G implies $i < j$.

A graph is *undirected* if it has no arrow. An undirected graph is called *chordal*, or *decomposable*, if every (undirected) cycle of length at least 4 has a chord, that is, an edge connecting two non-consecutive nodes in the cycle. There is a number of equivalent definitions for a decomposable graph (Lauritzen, 1996); one of them says that it is an undirected graph which can be acyclically directed without creating an immorality. A special case of a chordal graph is a *forest*, which is an undirected graph without undirected cycles. A forest over N in which N is connected is called a *(spanning) tree*.

A *chain graph* is a graph G (allowing both directed and undirected edges) whose components can be ordered into a chain, which is a sequence $C_1, \ldots, C_m$, $m \geq 1$ such that if $a \to b$ in G then $a \in C_i$ and $b \in C_j$ with $i < j$. An equivalent definition is: It is a graph without semi-directed cycles. Of course, every acyclic directed graph and every undirected graph is a special case of a chain graph (without flags).

Given a connected set C in a chain graph G, the set of *parents* of C is

$$pa_G(C) \equiv \{\, a \in N;\ a \to b \text{ in } G \text{ for some } b \in C \,\}.$$

Clearly, in a chain graph, $pa_G(C)$ is disjoint with (a connected set) C.

2.2 Bayesian network structures

Let N be a finite set of *variables*; to avoid the trivial case assume $|N| \geq 2$. For each $i \in N$ consider a finite individual *sample space* X_i (of possible values); to avoid technical problems assume $|\mathsf{X}_i| \geq 2$, for each $i \in N$. A *Bayesian network* can be introduced as a pair (G, P), where G is an acyclic directed graph having N as the set of its nodes and P a probability distribution on the *joint sample space* $\prod_{i \in N} \mathsf{X}_i$ that recursively factorizes according to G. Note that a factorization of P is equivalent to the condition that P is *Markovian* with respect to G meaning that it satisfies conditional independence restrictions determined by the respective separation criterion (Lauritzen, 1996).

BN structure (= Bayesian network structure) defined by an acyclic directed graph G is formally the class of probability distributions (on a fixed joint sample space) being Markovian with respect to G. Different graphs over N can be *Markov equivalent*, which means they define the same BN structure. The classic graphical characterization of (Markov) equivalent graphs is as follows (Verma and Pearl, 1991): they are equivalent if the have the same underlying undirected graph (= adjacencies) and the same immoralities. Of course, a BN structure can be described by any acyclic directed graph defining it, but there are other representatives (see below).

A complete *database* D of length $\ell \geq 1$ is a sequence $x_1, \ldots, x_\ell$ of elements of the joint sample space. By *learning BN structure* (from data) is meant to determine the BN structure based on an observed database D. A *quality criterion* is a real function $\mathcal{Q}$ of two variables: of an acyclic directed graph G and of a database D. The value $\mathcal{Q}(G, D)$ evaluates quantitatively how good the BN structure defined by G is to explain the occurrence of the database D. However, we will not repeat the formal definition of the relevant concept of *statistical consistency* of $\mathcal{Q}$; see (Neapolitan, 2004). Since the aim is to learn a BN structure, a natural requirement is $\mathcal{Q}$ to be *score equivalent*, i.e., for fixed D, we have

$$\mathcal{Q}(G, D) = \mathcal{Q}(H, D),$$

for any pair of Markov equivalent acyclic directed graphs G and H over N.

An additively *decomposable* criterion (Chickering, 2002) is a criterion which can be written as follows:

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

where D_A for $\emptyset \neq A \subseteq N$ is the projection of the database D to $\prod_{i \in A} \mathsf{X}_i$ and $q_{i|B}$ for $i \in N$, $B \subseteq N \setminus \{i\}$ are real functions.

Statistical scoring methods are typically based on the likelihood function. For example, evaluating each BN structure by a *maximized log-likelihood* (MLL) leads to a score equivalent and additively decomposable criterion. However, this criterion is not statistically consistent in the sense of (Neapolitan, 2004), because it does not take the complexity of statistical models into consideration. Therefore, subtracting a penalty term evaluating the dimension of the statistical model and the length of the database may solve the problem. A standard example of such a criterion which is statistically consistent, score equivalent and decomposable is Schwarz's *Bayesian information criterion* (BIC) (1978).

2.3 Essential graph

The *essential graph* G^* of an equivalence class $\mathcal{G}$ of acyclic directed graphs over N is defined as follows:

- $a \to b$ in G^* if $a \to b$ in every G from $\mathcal{G}$,
- a and b are adjacent by an undirected edge in G^* if there are graphs G_1 and G_2 in $\mathcal{G}$ such that $a \to b$ in G_1 and $a \leftarrow b$ in G_2.

The first graphical characterization of essential graphs was provided by Andersson, Madigan and Perlman (1997). It follows from this characterization that every essential graph is a chain graph without flags.

Actually, chain graphs without flags can serve as convenient graphical representatives of BN structures. As explained in Section 2.3 of (Studený, Roverato and Štěpánová, 2009), every chain graph defines a class of Markovian distributions, a statistical model, through the respective (generalized) separation criterion. As in case of acyclic directed graphs, they are called *Markov equivalent* if they define the same statistical model. Lemma 3 in (Studený, 2004) states that a chain graph H without flags is equivalent to an acyclic directed graph if the induced subgraphs for its components are chordal (undirected) graphs. Moreover, we can extend the graphical characterization of equivalence: two chain graphs without flags are Markov equivalent iff they have the same adjacencies and immoralities; see Lemma 2 in (Studený, 2004).

In this paper, we exploit the following characterization of essential graphs: Given an acyclic directed graph G, let $\mathcal{G}$ be the equivalence class of acyclic directed graphs containing G and $\mathcal{H}$ the (wider) equivalence class of chain graphs without flags containing G. The class $\mathcal{H}$ can be naturally (partially) ordered: if $H_1, H_2 \in \mathcal{H}$ and $a \to b$ in H_1 implies $a \to b$ in H_2 we call H_1 to be *larger* than H_2. With this partial ordering, the essential graph G^* (of $\mathcal{G}$) is just the largest graph in $\mathcal{H}$; see Corollary 4 in (Studený, 2004).

Moreover, there is a graphical procedure for getting G^* on the basis of any G in $\mathcal{G}$. It is based on a special graphical operation. Let H be a chain graph without flags. Consider two of its components, U called the *upper component* and L called the *lower component.* Provided the following two conditions hold:

- $pa_H(L) \cap U \neq \emptyset$ is a complete set in H,
- $pa_H(L) \setminus U = pa_H(U)$,

we say that the components can be *legally merged.* The result of merging is a graph obtained from H by replacing the arrows directed from U to L into undirected edges. By Corollary 26 in (Studený, Roverato and Štěpánová, 2009), the resulting graph is also a chain graph without flags equivalent to H. Moreover, Corollary 28 in (2009) says: If G and H are equivalent chain graphs without flags and H is larger than G, then there exists a sequence of legal merging operations which successively transforms G into H. Of course, this is applicable to an acyclic directed graph G and the essential graph G^* in place of H.

2.4 Algebraic approach

In this paper, we consider vectors whose components are ascribed to (= indexed by) subsets of the set of variables N. Let $\mathcal{P}(N) \equiv \{A;\ A \subseteq N\}$ denote the power set of N. Every element of $\mathbb{R}^{|\mathcal{P}(\mathcal{N})|}$ can be written as a (real) combination of basic imsets vectors $\delta_A \in \{0,1\}^{|\mathcal{P}(\mathcal{N})|}$, $A \subseteq N$, where $\delta_A(A) = 1$ and $\delta_A(B) = 0$ for $A \neq B \subseteq N$.

Given an acyclic directed graph G over N, the *standard imset* for G in $\mathbb{R}^{|\mathcal{P}(\mathcal{N})|}$ is defined by the formula

$$u_G = \delta_N - \delta_\emptyset + \sum_{i\in N} \left\{ \delta_{pa_G(i)} - \delta_{\{i\}\cup pa_G(i)} \right\}, \quad (2)$$

where the basic vectors can cancel each other. An important fact is that two acyclic directed graphs G and H over N are Markov equivalent iff $u_G = u_H$; see Corollary 7.1 in (Studený, 2005). The crucial fact, however, is: Every score equivalent and decomposable criterion $\mathcal{Q}$ has the form (1), where $s_D^{\mathcal{Q}} \in \mathbb{R}$ and $t_D^{\mathcal{Q}} \in \mathbb{R}^{|\mathcal{P}(\mathcal{N})|}$ only depend on the data (and $\mathcal{Q}$); see Lemmas 8.3 and 8.7 in (2005). Moreover, (the constant $s_D^{\mathcal{Q}}$ and) the *data vector* $t_D^{\mathcal{Q}}$ is uniquely determined under additional standardization conditions $t_D^{\mathcal{Q}}(A) = 0$ for $A \subseteq N$ with $|A| \leq 1$.

For example, the standardized data vector for the MLL criterion can be computed as follows; see Proposition 8.4 in (2005). Let $\hat{P}$ denote the empirical measure on $\prod_{i\in N} \mathsf{X}_i$ computed from D and $\hat{P}_A$ its marginal for $A \subseteq N$. The *multiinformation* of $\hat{P}_A$ (for $A \neq \emptyset$) is its relative entropy $H(\hat{P}_A | \prod_{i\in A} \hat{P}_{\{i\}})$ with respect to the product of its own one-dimensional marginals. Then $t_D^{\mathsf{MLL}}(A) = \ell \cdot H(\hat{P}_A | \prod_{i\in A} \hat{P}_{\{i\}})$, where ℓ is the length of the database D. A formula for the data vector relative to the BIC criterion can be found in Section 8.4.2 of (Studený, 2005).

3 Characteristic imset

The characteristic imset is formally an element of $\mathbb{Z}^{|\mathcal{P}_*(\mathcal{N})|}$, where $\mathcal{P}_*(N) \equiv \{A \subseteq N;\ |A| \geq 2\}$ is the class of sets of cardinality at least 2.

Definition 1. Given an acyclic directed graph G over N, the *characteristic imset* for G is given by the formula

$$c_G(A) = 1 - \sum_{B,\, A\subseteq B\subseteq N} u_G(B)\,, \quad (3)$$

for $A \subseteq N$, $|A| \geq 2$.

Clearly, the characteristic imset is obtained from the standard one by an affine transformation of $\mathbb{R}^{|\mathcal{P}(\mathcal{N})|}$ to $\mathbb{R}^{|\mathcal{P}_*(\mathcal{N})|}$ (we only add and subtract entries of u_G). This mapping is invertible: We can compute back the standard imset by the formula

$$u_G(B) = \sum_{A,\, B\subseteq A\subseteq N} (-1)^{|A\setminus B|} \cdot (1 - c_G(A)) \quad (4)$$

for $B \subseteq N$, $|B| \geq 2$. The remaining values of u_G can then be determined by the formulas $\sum_{S\subseteq N} u_G(S) = 0$ and $\sum_{S,i\in S\subseteq N} u_G(S) = 0$ for $i \in N$. Since the transformation is one-to-one, two acyclic directed graphs G and H are equivalent iff $c_G = c_H$ (cf. Section 2.4). Thus, the characteristic imset is also a unique BN structure representative.

The basic observation is as follows; see also Theorem 3.2 in (Hemmecke et al., 2010):

Theorem 1. *For any acyclic directed graph G over N we have $c_G(A) \in \{0,1\}$ for any $A \subseteq N$, $|A| \geq 2$. Moreover, $c_G(A) = 1$ iff there exists $i \in A$ with $A \setminus \{i\} \subseteq pa_G(i)$.*

Proof. First, we substitute (2) into (3) and get for fixed $A \subseteq N$, $|A| \geq 2$:

$$\begin{aligned} c_G(A) &= -\sum_{i\in N, A\subseteq pa_G(i)} 1 + \sum_{i\in N, A\subseteq \{i\}\cup pa_G(i)} 1 \\ &= \sum_{i\in N, A\subseteq \{i\}\cup pa_G(i)\ \&\ i\in A} 1 = \sum_{i\in A, A\setminus\{i\}\subseteq pa_G(i)} 1\,. \end{aligned}$$

Assume for a contradiction there exist distinct $i, j \in A$ with $A \setminus \{i\} \subseteq pa_G(i)$ and $A \setminus \{j\} \subseteq pa_G(j)$. Then, however, both $j \to i$ and $i \to j$ are in G contradicting its acyclicity. In particular, $c_G(A) \in \{0, 1\}$. □

The consequence is the characterization of adjacencies and immoralities in terms of the characteristic imset.

Corollary 1. *Let G be an acyclic directed graph over N and a, b (and c) are distinct nodes. Then*

(i) *a and b are adjacent in G iff $c_G(\{a, b\}) = 1$.*

(ii) *$a \to c \leftarrow b$ is an immorality in G iff $c_G(\{a, b, c\}) = 1$ and $c_G(\{a, b\}) = 0$. The latter two conditions imply $c_G(\{a, c\}) = 1$ and $c_G(\{b, c\}) = 1$.*

Proof. Part (i) directly follows from Theorem 1: $c_G(\{a, b\}) = 1$ iff either $b \in pa_G(a)$ or $a \in pa_G(b)$. The necessity of the condition in (ii) also follows from Theorem 1. Conversely, if $c_G(\{a, b, c\}) = 1$, three options may occur: $\{b, c\} \subseteq pa_G(a)$, $\{a, c\} \subseteq pa_G(b)$ and $\{a, b\} \subseteq pa_G(c)$. But $c_G(\{a, b\}) = 0$ means by (i) that a and b are not adjacent in G, which excludes the first two options and implies that $a \to c \leftarrow b$ is an immorality in G. □

Now we show that any reasonable quality criteria is an affine function of the characteristic imset.

Definition 2. Given a score equivalent, additively decomposable criterion $\mathcal{Q}$ and a database D, let $t_D^{\mathcal{Q}}$ denote the standardized data vector relative to $\mathcal{Q}$. Introduce the *revised data vector* (relative to $\mathcal{Q}$) as an element of $\mathbb{R}^{|\mathcal{P}_*(\mathcal{N})|}$:

$$r_D^{\mathcal{Q}}(A) = \sum_{B, B \subseteq A, |B| \geq 2} (-1)^{|A \setminus B|} \cdot t_D^{\mathcal{Q}}(B) \qquad (5)$$

for $A \subseteq N$, $|A| \geq 2$.

Lemma 1. *Every score equivalent and additively decomposable criterion $\mathcal{Q}$ has the form*

$$\mathcal{Q}(G, D) \;=\; \mathcal{Q}(G^{\emptyset}, D) + \langle r_D^{\mathcal{Q}}, c_G \rangle \,, \qquad (6)$$

where $G^{\emptyset}$ is the graph over N without edges.

Proof. We substitute (4) into (1):

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \sum_{B \subseteq N, |B| \geq 2} t_D^{\mathcal{Q}}(B) \cdot \overbrace{\sum_{A, B \subseteq A} (-1)^{|A \setminus B|} \cdot (1 - c_G(A))}^{u_G(B)} \,.$$

Now, change the order of summation in the sum:

$$\sum_{A \subseteq N, |A| \geq 2} (1 - c_G(A)) \cdot \underbrace{\sum_{B \subseteq A, |B| \geq 2} (-1)^{|A \setminus B|} \cdot t_D^{\mathcal{Q}}(B)}_{r_D^{\mathcal{Q}}(A)} \,.$$

Thus, we get by (5):

$$\begin{aligned} \mathcal{Q}(G, D) &= s_D^{\mathcal{Q}} - \sum_{A \subseteq N, |A| \geq 2} (1 - c_G(A)) \cdot r_D^{\mathcal{Q}}(A) \\ &= \textit{constant} + \sum_{A \subseteq N, |A| \geq 2} c_G(A) \cdot r_D^{\mathcal{Q}}(A) \,. \end{aligned}$$

The observation that the characteristic imset for the empty graph $G^{\emptyset}$ is identically zero implies that the *constant* above is simply $\mathcal{Q}(G^{\emptyset}, D)$. □

Finally, we establish the relation of the characteristic imset to any chain graph without flags defining the BN structure.

Theorem 2. *Let H be a chain graph without flags equivalent to an acyclic directed graph G. For any $A \subseteq N$, $|A| \geq 2$ one has $c_G(A) = 1$ iff*

$$\exists\, \emptyset \neq K \subseteq A \textit{ complete in } H, \textit{ with } A \setminus K \subseteq pa_H(K). \quad (7)$$

Proof. In an acyclic directed graph G, the only non-empty complete sets are singletons. Thus, by Theorem 1, $c_G(A) = 1$ iff (7) holds with G (in place of H).

The next step is to observe that if $\tilde{H}$ is obtained from a chain graph H without flags by legal merging of components (see Section 2.3), then for any $A \subseteq N$, $|A| \geq 2$, (7) holds with H iff it holds with $\tilde{H}$. To verify this observe that any set A satisfying (7) has a uniquely determined component C with $K \subseteq C$ in H. Moreover, $pa_H(K) = pa_H(C)$, since H has no flags. The validity of (7) then depends on the induced subgraph of H for $C \cup pa_H(C)$. However, if $\tilde{H}$ is obtained from H by legal component merging, then most of these induced subgraphs are kept and the only change concerns the merged components U and L. We leave the reader to evidence that this change satisfies condition (7) in both directions.

Finally, we use the result mentioned in Section 2.3 which implies the existence of sequences of legal merging operations transforming G into G^* and H into G^*. In particular, for $A \subseteq N$, $|A| \geq 2$, (7) with G is equivalent to (7) with G^*, and this is equivalent to (7) with H. □

Of course, Theorem 2 applied to the essential graph G^* in place of H gives a direct method for obtaining the characteristic imset from the essential graph.

4 Back to the essential graph

Corollary 1 allows us to reconstruct the essential graph from the characteristic imset. Indeed, conditions (i) and (ii) determine both the adjacencies and immoralities (in every acyclic directed graph G defining the corresponding BN structure). Thus, we can directly get the *pattern* (of G) being the underlying undirected graph in which only the edges belonging to an immorality are directed.

This graph neither has to be the essential graph nor a chain graph. However, there is a simple (polynomial-time) procedure for transforming the pattern into the corresponding essential graph G^*. It consists of an (repeated) application of three orientation rules. Specifically, Theorem 3 in (Meek, 1995) states that the exhaustive application of rules from Figure 1 to the pattern of an acyclic directed graph G results in the essential graph (of the equivalence class containing G).

5 Learning undirected forests

Decomposable models (Lauritzen, 1996) can be viewed as BN structures whose essential graphs are (chordal) undirected graphs.

Corollary 2. *Let H be a chordal undirected graph over N. Then the corresponding characteristic imset c_H is specified as follows: $c_H(A) = 1$ iff A is a complete set in H.*

Proof. Consider the equivalence class $\mathcal{G}$ of acyclic directed graphs equivalent to H and apply Theorem 2. Since H has no arrow, (7) is equivalent to the above requirement. □

A special case of a chordal graph is an undirected forest. The only complete sets of cardinality at least 2 in it are its edges:

Corollary 3. *Let H be an undirected forest. Then the corresponding characteristic imset c_H vanishes for sets of cardinality* 3 *and more, and for distinct $a, b \in N$ we have $c_H(\{a, b\}) = 1$ iff a and b are adjacent in H.*

In particular, the characteristic imset for a forest can be identified with a vector of polynomial length $\binom{|N|}{2}$, which simplifies many things. For example, if maximizing a quality criterion $\mathcal{Q}$ over (undirected) forests is of interest, then, by Lemma 1, the function $H \mapsto c_H \in \mathbb{Z}^{|\mathcal{P}_*(\mathcal{N})|} \mapsto \langle r_D^{\mathcal{Q}}, c_H \rangle = \sum_{A \text{ edge in } H} r_D^{\mathcal{Q}}(A)$ should be maximized, that is, $H \mapsto \sum_{A \text{ edge in } H} t_D^{\mathcal{Q}}(A)$ by (5).

In particular, in case of the MLL criterion this means maximizing the sum of weights $\sum_{\{a,b\} \text{ edge}} w_{\{a,b\}}$, where $w_{\{a,b\}} = H(\hat{P}_{\{a,b\}} | \hat{P}_{\{a\}} \times \hat{P}_{\{b\}})$ is the (empirical) *mutual information* between a and b; see Section 2.4.

The polytope spanned by (restricted) characteristic imsets for forests has already been studied in matroid theory (Schrijver, 2003). It appears to be quite nice from an algorithmic point of view – for details see (Hemmecke et al., 2010). One important observation is the existence of a simple polynomial-time procedure based on the *greedy algorithm* for finding maximum-weight forest, where forests are weighted by the sums of weights of their edges.

This gives an elegant geometric interpretation to a classic (heuristic) procedure for approximating probability distributions with trees proposed by Chow and Liu (1968). Taking into account what was said above, it can be interpreted as the maximization of the MLL criterion over trees (= connected forests) using the greedy technique.

Conclusions

Our geometric interpretation of the classic learning procedure (for trees) may lead to useful generalizations. First, the application of the greedy algorithm is not limited to the MLL criterion and can be applied to maximize other

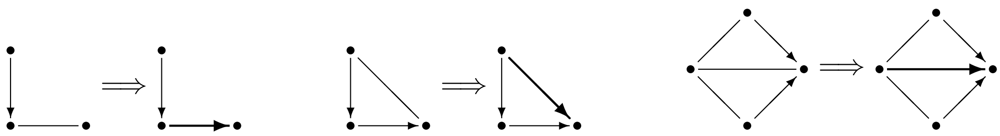

Figure 1: Orientation rules for getting the essential graph.

reasonable criteria like the BIC criterion. Second, we are not limited to trees and can apply the method to learning undirected forests, actually, to learning sub-forests of a prescribed undirected graph. Future research topics could be whether characteristic imsets can be applied to learning decomposable models, for example, with limited cardinality of cliques.

There are some related open questions. It follows from Section 4 that the components of the characteristic imset for sets of cardinalities 2 and 3 determine the remaining components. However, is there any direct method for determining them? Another question is whether Meek's (1995) orientation rules can be avoided in the reconstruction of the essential graph on the basis of the characteristic imset. We hope that a modification of the procedure from (Studený and Vomlel, 2009) leads to such an algorithm.

Acknowledgments

This research has been supported by the grants GAČR n. 201/08/0539 and MŠMT n. 1M0572.

References

Steen A. Andersson, David Madigan and Michael D. Perlman. 1997. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25(2):505–541.

Remco R. Bouckaert. 1995. Bayesian belief networks: from construction to evidence. PhD thesis, University of Utrecht.

David M. Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

C. K. Chow, C. N. Liu. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Raymond Hemmecke, Silvia Lindner, Milan Studený, and Jiří Vomlel. 2010. Characteristic imsets for learning Bayesian network structures. In preparation.

Steffen L. Lauritzen. 1996. *Graphical Models*, Clarendon Press.

Chris Meek. 1995. Causal inference and causal explanation with background knowledge. In *11th Conference on Uncertainty in Artificial Intelligence*, pages 403–410.

Richard E. Neapolitan. 2004. *Learning Bayesian Networks*, Pearson Prentice Hall.

Alexander Schrijver. 2003. *Combinatorial Optimization - Polyhedra and Efficiency, volume B*, Springer Verlag.

Gideon E. Schwarz. 1978. Estimation of the dimension of a model. *Annals of Statistics*, 6:461–464.

Milan Studený. 2004. Characterization of essential graphs by means to the operation of legal merging of components. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12:43–62.

Milan Studený. 2005. *Probabilistic Conditional Independence Structures*, Springer Verlag.

Milan Studený and Jiří Vomlel. 2009. A reconstruction algorithm for the essential graph. *International Journal of Approximate Reasoning*, 50:385–413.

Milan Studený, Alberto Roverato and Šárka Štěpánová. 2009. Two operations of merging and splitting components in a chain graph. *Kybernetika*, 45(2):208–248.

Milan Studený, Jiří Vomlel and Raymond Hemmecke. 2010. A geometric view on learning Bayesian network structures. *International Journal of Approximate Reasoning*, 51(5):578–586.

Thomas Verma and Judea Pearl. 1991. Equivalence and synthesis of causal models. In *6th Conference on Uncertainty in Artificial Intelligence*, pages 220–227.