International Journal of Approximate Reasoning 53 (2012) 1336–1349

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Characteristic imsets for learning Bayesian network structure

Raymond Hemmecke [a], Silvia Lindner [a], Milan Studený [b,*]

[a] Zentrum Mathematik, Technische Universität München, Munich, Germany
[b] Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic

ARTICLE INFO

Article history:
Available online 6 May 2012

Keywords:
Learning Bayesian network structure
Essential graph
Standard imset
Characteristic imset
LP relaxation of a polytope

ABSTRACT

The motivation for the paper is the geometric approach to learning Bayesian network (BN) structure. The basic idea of our approach is to represent every BN structure by a certain uniquely determined vector so that usual scores for learning BN structure become affine functions of the vector representative. The original proposal from Studený et al. (2010) [26] was to use a special vector having integers as components, called the *standard imset*, as the representative. In this paper we introduce a new unique vector representative, called the *characteristic imset*, obtained from the standard imset by an affine transformation.

Characteristic imsets are (shown to be) zero-one vectors and have many elegant properties, suitable for intended application of linear/integer programming methods to learning BN structure. They are much closer to the graphical description; we describe a simple transition between the characteristic imset and the *essential graph*, known as a traditional unique graphical representative of the BN structure. In the end, we relate our proposal to other recent approaches which apply linear programming methods in probabilistic reasoning.

1. Introduction

Bayesian networks are basic graphical models, used widely both in statistics [13] and artificial intelligence [18]. These statistical models of *conditional independence* (CI) structure are described by acyclic directed graphs whose nodes correspond to (random) variables in consideration. It may happen that two different graphs describe the same statistical model, that is, they are *Markov equivalent*. A classic result [10,31] says that two acyclic directed graphs are Markov equivalent iff they have the same adjacencies and *immoralities*, which are special induced subgraphs over three nodes.

A quite important topic is learning *Bayesian network* (BN) *structure* [16], which is determining the statistical model on the basis of observed data. Although there are learning methods based on statistical CI tests, contemporary score and search methods are based on maximization of a suitable *quality criterion* $\mathcal{Q}$, also named a *scoring criterion* or simply a *score* by some 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 assumptions 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* [3] and (additively) *decomposable* [5].

Representing the 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 particular BN structure. A classic unique graphical representative is the *essential graph* [1] of the corresponding Markov equivalence class of acyclic directed graphs, which is a special graph allowing both directed and undirected edges.

* Corresponding author.
E-mail addresses: hemmecke@ma.tum.de (R. Hemmecke), slindner@ma.tum.de (S. Lindner), studeny@utia.cas.cz (M. Studený).

0888-613X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2012.04.001

The basic idea of an algebraic approach to the description of CI structures [22] is to represent them by certain vectors with integer components, called *imsets*. In the context of learning Bayesian networks this led to the proposal to represent each BN structure uniquely by a so-called *standard imset*. The advantage of this algebraic approach is that every score equivalent and decomposable quality criterion becomes an affine function (= linear function plus a constant) of the standard imset (see Chapter 8 in [22]).

A geometric view was offered in [26], where it was shown that the standard imsets over a fixed set of variables N are vertices (= extreme points) of a certain polytope, called the *standard imset polytope*. These results allow one to use the tools of polyhedral geometry in the area of learning Bayesian nets, because they transform the learning task to a *linear programming* (LP) problem [19], namely to optimize a linear function over a bounded polyhedron.

In this paper, we propose an alternative vector representative of the BN structure, called the *characteristic imset*. It is a vector obtained from the standard imset by a one-to-one affine transformation which maps integer vectors to integer vectors (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, called the *characteristic imset polytope*. Every characteristic imset is a zero-one vector, that is, it has only zeros and ones as its components. Moreover, it is very close to the graphical description: both adjacencies and immoralities are encoded by certain components of the characteristic imset (see Corollary 2 for details).

We establish a simple relation of the characteristic imset to any acyclic directed graph defining the BN structure and to the respective essential graph as well. More specifically, we provide a formula for the characteristic imset on basis of any chain graph defining the BN structure which has no *flag*, which is a special induced subgraph over three nodes. In particular, this makes it possible to get the characteristic imset immediately on the basis of the essential graph. We also consider the converse task of reconstructing the essential graph from the characteristic imset and provide a polynomial algorithm (in the number $|N|$ of variables) for it.

If we restrict our attention to *decomposable models* [13], interpreted as BN structures, then the characteristic imset has quite a simple form. The situation is particularly transparent in the case of (models induced by) *undirected forests*: then the edges in the graph correspond to ones in the characteristic imset. Thus, one can use the well-known *greedy algorithm* [20] to learn these special graphical models; this gives an elegant geometric interpretation to the classic heuristic procedure proposed by Chow and Liu [6].

The structure of the paper, which is based on [27], 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 transition between the essential graph and the characteristic imset. Section 5 contains comments on learning decomposable models. In Section 6 we relate characteristic imsets to other zero-one vector structure representatives which have recently appeared in the literature [17,11]. We also discuss there the idea of intended future application of this approach to practical learning, motivated by [11,7,9]. In Section 7 we briefly mention our preliminary computational experiments, and, in Conclusions we discuss the perspectives.

2. Basic concepts

We tacitly assume that the reader is familiar with basic concepts from polyhedral geometry. Throughout the paper N is a finite non-empty set of *variables*; to avoid the trivial case we assume $|N| \geq 2$. In statistical context, the elements of N correspond to random variables in consideration; in graphical context, they correspond to nodes.

2.1. Graphical concepts

Graphs considered here 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 loops or multiple edges are allowed between two nodes. If there is an edge between nodes i and j, we say they are *adjacent*.

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. To avoid confusion note that some authors [13] may use this term to name a set in which every pair of distinct nodes is adjacent, no matter whether by a directed or an undirected edge. However, in this paper we do need the stronger concept of (an undirected) complete set. 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 (undirected) *components*. Of course, the components of G provide a natural partition of N.

A graph is *directed* if all its edges are arrows. A directed graph G over N is called *acyclic* if there exists an ordering $b_1, \ldots, b_{|N|}$ of all its nodes 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 all its edges are undirected. An undirected graph is called *chordal*, or *decomposable*, if every (undirected) cycle of the length at least four has a chord, that is, an edge connecting two non-consecutive nodes in the cycle.

There is a number of equivalent definitions of a decomposable graph [13]; 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 a connected set is called a (*spanning*) *tree*.

A *chain graph* is a graph whose (undirected) 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 $i < j$. Every acyclic directed graph and every undirected graph is a special case of a chain graph; in fact, they both fall in the class of chain graphs without flags, which class of graphs plays an important role below.

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

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

Clearly, in a chain graph, if C is a connected set of nodes in G, then $pa_G(C)$ is disjoint with C. Typically, we will have a singleton $\{i\}$ in place of C, in which case we write $pa_G(i)$.

2.2. Learning Bayesian network structure

In a statistical context, each variable (= node) $i \in N$ is assigned a finite (individual) sample space X_i (= the set of possible values); to avoid technical problems assume $|\mathsf{X}_i| \geq 2$ for each $i \in N$. A *BN structure* defined by an acyclic directed graph G (over N) is formally the class of discrete probability distributions P on the joint sample space $\prod_{i\in N} \mathsf{X}_i$ that are Markovian with respect to G. Recall from [13,18] that P is *Markovian* with respect to G if it satisfies CI restrictions determined by the respective separation criterion.

As mentioned in the Introduction a classic graphical equivalence characterization says that two acyclic directed graphs over N are *Markov equivalent* (= define the same BN structure) iff they have the same adjacencies and immoralities (for a proof see [1]). The *inclusion* of BN structures is the inclusion of corresponding classes of probability distributions.

A complete *database* D of the 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 determining 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 $\mathcal{Q}$; for details see [16].

Since the aim is to learn a BN structure, a natural requirement is $\mathcal{Q}$ to be *score equivalent* [3], that is, 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 [5] is a criterion $\mathcal{Q}$ which can be written as follows:

$$\mathcal{Q}(G, D) = \sum_{i\in N} q_{i|pa_G(i)}(D_{\{i\}\cup pa_G(i)}), \tag{1}$$

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. Note that the terms $q_{i|B}(D_{\{i\}\cup B})$ are often called the *local scores*.

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 sense of [16], because it does not take into consideration the complexity of statistical models. 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) [21].

Bayesian approach to derive criteria is to average the likelihood function after some prior distributions on the respective parameter spaces; then each BN structure can be evaluated by the *logarithm of the marginal likelihood*. However, a plenty of technical assumptions must be accepted here to make it consistent [12]; an example of a score equivalent and decomposable criterion derived in this way is the *Bayesian Dirichlet Equivalence* (BDE) score.

2.3. Essential graph

The *essential graph* G^* of a Markov 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 et al. [1]. It follows from that characterization that every essential graph is a chain graph without flags.

In this paper, we exploit the following characterization of essential graphs from §6.5 of [25]. 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. Here, we consider two chain graphs over N without flags equivalent if they have the same adjacencies and immoralities. The class $\mathcal{H}$ can be naturally (partially) ordered as follows:

$$\text{given } H_1, H_2 \in \mathcal{H}\,, \qquad \text{if, } \forall\, a, b \in N\,, \quad a \to b \text{ in } H_1 \text{ implies } a \to b \text{ in } H_2,$$

then we say that H_1 is *larger* than H_2. This terminology was introduced by Frydenberg [10] and the reason was as follows. If H_1 is larger than H_2 then this implies (but is not equivalent to!) that any undirected edge in H_2 is an undirected edge in H_1, meaning that H_1 is more "undirected" than H_2. With this partial ordering, the essential graph G^* (of $\mathcal{G}$) is just the largest graph in $\mathcal{H}$; see Proposition 29 in [25].

Moreover, there is a graphical procedure for getting G^* on the basis of any graph H in $\mathcal{H}$. 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 by undirected edges. By Corollary 26 in [25], the resulting graph is also a chain graph without flags equivalent to H. Moreover, Corollary 28 in [25] 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

An *imset* over N is a vector, whose components are integers indexed by subsets of N. Traditionally, all subsets of N are considered, although in Section 3 we also consider imsets with a restricted domain. Thus, in the terminology of polyhedral geometry, imsets are just the *lattice points* in the Euclidean space $\mathbb{R}^{\mathcal{P}(N)}$ (= elements of $\mathbb{Z}^{\mathcal{P}(N)}$), where $\mathcal{P}(N) \equiv \{A;\ A \subseteq N\}$ denotes the power set of N.

Every vector in this space can be written as a (real) combination of basic vectors $\delta_A \in \{0, 1\}^{\mathcal{P}(N)}$ for $A \subseteq N$:

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

This allows us to write formulas for imsets. Given an acyclic directed graph G over N, the *standard imset* for G is given by

$$\mathrm{u}_G := \delta_N - \delta_\emptyset + \sum_{i \in N} \left\{ \delta_{pa_G(i)} - \delta_{\{i\} \cup pa_G(i)} \right\}, \tag{2}$$

where the basic vectors can cancel each other. It is a unique algebraic representative of the corresponding BN structure because $\mathrm{u}_G = \mathrm{u}_H$ if and only if G and H are Markov equivalent (Corollary 7.1 in [22]). The convex hull of the set of all standard imsets over N is called the *standard imset polytope*.

Although the standard imset is a vector of an exponential length in $|N|$, the memory demands for its computer representation are polynomial in $|N|$. This is because at most $2 \cdot |N|$ of its components are non-zero, since at least one term $\delta_{pa_G(i)}$ in (2) cancels against $-\delta_\emptyset$. Note that there is a polynomial-time algorithm (in $|N|$) for the reconstruction of the essential graph from the standard imset [24].

2.4.1. Algebraic view on learning

An important result from the point of view of an algebraic approach to learning BN structure is that any score equivalent and decomposable criterion $\mathcal{Q}$ is an affine function of the standard imset. Specifically, $\mathcal{Q}$ has the form

$$\mathcal{Q}(G, D) \;=\; s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, \mathrm{u}_G \rangle\,, \tag{3}$$

where $\langle *, * \rangle$ denotes the scalar product, and both $s_D^{\mathcal{Q}} \in \mathbb{R}$ and $t_D^{\mathcal{Q}} \in \mathbb{R}^{\mathcal{P}(N)}$ only depend on the database D and the criterion (see Lemmas 8.3 and 8.7 in [22]). In particular, the task to maximize $\mathcal{Q}$ is equivalent to finding the optimum of a linear function over the standard imset polytope. 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 $|A| \leq 1$.

Note that one can hardly expect that most of the components of the (standardized) data vector are zeros. However, because the components of the standard imset u_G mostly vanish, to compute the value of $\mathcal{Q}(G, D)$ by means of (3), one only

needs to compute at most $2 \cdot |N|$ components of $t_D^{\mathcal{Q}}$. Thus, the idea was to have a formula for the data vector and to compute a particular component of $t_D^{\mathcal{Q}}$ only when needed.

For example, the standardized data vector t_D^{MLL} for the MLL criterion can be computed as follows (see Proposition 8.4 in [22]). Let $\hat{P}$ denote the empirical measure on $\prod_{i\in N} \mathsf{X}_i$ computed from D, $\hat{P}_A$ its marginal for $A \subseteq N$ and $\mathcal{H}(\hat{P}_A | \prod_{i\in A} \hat{P}_{\{i\}})$ the relative entropy of $\hat{P}_A$ with respect to the product of its own one-dimensional marginals. Then

$$t_D^{\mathrm{MLL}}(A) = \ell \cdot \mathcal{H}\left(\hat{P}_A | \prod_{i\in A} \hat{P}_{\{i\}}\right) \quad \text{where } \ell \text{ is the length of the database } D, \qquad \text{for any } A \subseteq N.$$

A formula for the data vector relative to the BIC criterion can be found in Section 8.4.2 of [22]; an analogous formula in the case of BDE score in §8.3 of [23].

The reader may raise doubts whether the computation of components of the data vector is treatable for practical use, or, whether one can keep such a long vector in the memory of a computer (if needed). As explained in Section 6.4, one can easily compute the values of the data vector from local scores. Thus, once one is able to compute the local scores, one should be able to compute the components of the data vector as well. As concerns the problem of memory demands, realize that the memory demands for keeping all components of a data vector are smaller than the memory demands for representing the database in the form of a (contingency) table of counts. Therefore, once one is able to keep the data in memory of a computer, one should be able to keep the data vector, too. Moreover, one can often in practice avoid the need for computing all components of the data vector (see Section 6.4).

2.4.2. CI statement coding

In the context of an algebraic description of CI structures [22], special simple imsets are used to describe CI statements. More specifically, given a triplet of pairwise disjoint sets of variables $A, B, C \subseteq N$, the corresponding *semi-elementary imset* is given by

$$\mathsf{u}_{\langle A,B|C\rangle} := \delta_{A\cup B\cup C} + \delta_C - \delta_{A\cup C} - \delta_{B\cup C}\,.$$

It algebraically encodes the CI statement $A \perp\!\!\!\perp B \,|\, C$, meaning that A is conditionally independent of B given C. The CI statement and the corresponding imset is called *elementary* if A and B are singletons: $|A| = |B| = 1$. Note that every semi-elementary imset is a standard one. The corresponding acyclic directed graph G over N can be obtained as follows. Order the variables in N in such a way that (the elements of) C precede A, then B follows and $N \setminus (A \cup B \cup C)$ is put at the end. Then direct the edges of the complete (undirected) graph over N according to this order and remove the arrows from A to B.

3. Characteristic imset

The characteristic imset is formally an imset with a restricted domain to the class

$$\mathcal{P}_2(N) := \{A \subseteq N;\ |A| \geq 2\} \quad \text{of sets of cardinality at least two.}$$

Definition 1. Given an acyclic directed graph G over N, let u_G be the standard imset for G. We introduce the (*upper*) *portrait* of u_G, denoted by p_G, as follows:

$$\mathsf{p}_G(S) := \sum_{T,\, S\subseteq T\subseteq N} \mathsf{u}_G(T) \qquad \text{for } S \subseteq N. \tag{4}$$

Then we subtract the portrait from the constant one-vector and get the *characteristic imset* for G, denoted by c_G:

$$\mathsf{c}_G(S) := 1 - \mathsf{p}_G(S) = 1 - \sum_{T,\, S\subseteq T\subseteq N} \mathsf{u}_G(T) \qquad \text{for } S \subseteq N. \tag{5}$$

We will consider c_G as an element of $\mathbb{Z}^{\mathcal{P}_2(N)}$, or equivalently, as an element of $\mathbb{Z}^{\mathcal{P}(N)}$ satisfying additional constraints $\mathsf{c}_G(S) = 1$ for $|S| \leq 1$. Having fixed N, the convex hull of the set of characteristic imsets is called the *characteristic imset polytope*.

The reason for ignoring the components of c_G for $|S| \leq 1$ is as follows. Every standard imset satisfies the linear constraints

$$\sum_{T\subseteq N} \mathsf{u}_G(T) = 0, \qquad \sum_{T,\, i\in T\subseteq N} \mathsf{u}_G(T) = 0 \quad \text{for any } i \in N. \tag{6}$$

In particular, one always has $\mathsf{p}_G(S) = 0$ for $S \subseteq N$ with $|S| \leq 1$, and therefore, $\mathsf{c}_G(S) = 1$ for those $S \subseteq N$.

Thus, the characteristic imset is obtained from the standard one by an invertible affine transformation of $\mathbb{R}^{\mathcal{P}(N)}$ to $\mathbb{R}^{\mathcal{P}_2(N)}$. Indeed, one can compute back the standard imset by the well-known formula for Möbius inversion:

$$u_G(T) = \sum_{S,\, T\subseteq S\subseteq N} (-1)^{|S\setminus T|}\cdot p_G(S) = \sum_{S,\, T\subseteq S\subseteq N} (-1)^{|S\setminus T|}\cdot (1 - c_G(S)) \quad \text{for } T\subseteq N, |T|\geq 2. \tag{7}$$

The remaining values of u_G can then be determined by (6). Since the transformation is one-to-one, two acyclic directed graphs G and H are Markov equivalent iff $c_G = c_H$ (cf. Section 2.4). Thus, the characteristic imset is also a unique BN structure representative.

The basic observation about the characteristic imset is as follows.

Theorem 1. *Let G be an acyclic directed graph over N. For any $\emptyset \neq S \subseteq N$ we have $c_G(S) \in \{0, 1\}$ and $c_G(S) = 1$ iff there exists some $i \in S$ with $S \setminus \{i\} \subseteq pa_G(i)$. In particular, c_G is a zero-one vector.*

Proof. Consider the defining formula (2) for the standard imset. For any $S \subseteq N$, $|S| \geq 1$, the entry $p_G(S)$ can be computed as

$$p_G(S) = 1 + \sum_{i\in N,\, S\subseteq pa_G(i)} 1 - \sum_{i\in N,\, S\subseteq\{i\}\cup pa_G(i)} 1\,.$$

Hence, we get

$$c_G(S) = 1 - p_G(S) = \sum_{i\in N,\, S\subseteq\{i\}\cup pa_G(i)} 1 - \sum_{i\in N,\, S\subseteq pa_G(i)} 1 = \sum_{i\in N,\, S\subseteq\{i\}\cup pa_G(i),\, i\in S} 1 = \sum_{i\in S,\, S\setminus\{i\}\subseteq pa_G(i)} 1\,.$$

For fixed S, assume that there exists two different elements $i, j \in S$ with $S \setminus \{i\} \subseteq pa_G(i)$ and $S \setminus \{j\} \subseteq pa_G(j)$. This implies both $i \in pa_G(j)$ and $j \in pa_G(i)$. The simultaneous existence of the arrows $i \to j$ and $j \to i$, however, contradicts the fact of G being acyclic. Thus, for each $S \subseteq N$, there is at most one $i \in S$ with $S \setminus \{i\} \subseteq pa_G(i)$. Consequently,

$$c_G(S) = \sum_{i\in S,\, S\setminus\{i\}\subseteq pa_G(i)} 1 \quad \in \{0, 1\},$$

and thus c_G is a zero-one vector. □

Corollary 1. *For any N, the only lattice points in the characteristic imset polytope and in the standard imset polytope are their vertices.*

Proof. The result holds for any zero-one polytope and thus also for the characteristic imset polytope. The portrait map is an affine linear map between u_G and c_G, mapping lattice points to lattice points, in both directions. Thus, the result holds also for the standard imset polytope. □

Remark. Note that the observation that there is no lattice point in the standard imset polytope except its vertices is also made in [28]. The original proof of this result in the manuscript of [28] was quite long and complicated. However, later discussion among the authors of the present paper lead to a much simpler proof, namely using the portrait map. In the end, this simple proof, the result of our joint effort, was also used in the final version of [28].

Another consequence of Theorem 1 is the characterization of adjacencies and immoralities in terms of the characteristic imset.

Corollary 2. *Let G be an acyclic directed graph over N and a, b (and c) are distinct nodes in G. 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. □

3.1. Quality criteria and characteristic imsets

Now we show that any usual quality criterion is an affine function of the characteristic imset.

Definition 2. Given a score equivalent, an 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 $\mathbb{R}^{\mathcal{P}_2(N)}$:

$$r_D^{\mathcal{Q}}(A) = \sum_{B, B\subseteq A, |B|\geq 2} (-1)^{|A\setminus B|} \cdot t_D^{\mathcal{Q}}(B) \qquad \text{for } A \subseteq N, |A| \geq 2. \tag{8}$$

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

$$\mathcal{Q}(G, D) \;=\; \mathcal{Q}(G^{\emptyset}, D) + \langle r_D^{\mathcal{Q}}, \mathsf{c}_G \rangle \,, \tag{9}$$

where $G^{\emptyset}$ is the empty graph over N (= graph without adjacencies).

Proof. Realize that $t_D^{\mathcal{Q}}(B) = 0$ for $|B| \leq 1$ and substitute (7) into (3):

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \sum_{B\subseteq N,\ |B|\geq 2} t_D^{\mathcal{Q}}(B) \cdot \underbrace{\sum_{A,\ B\subseteq A} (-1)^{|A\setminus B|} \cdot (1 - \mathsf{c}_G(A))}_{\mathsf{u}_G(B)} \,.$$

Now, change the order of summation in the latter sum:

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

Thus, we get by (8):

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \sum_{A\subseteq N,\ |A|\geq 2} (1 - \mathsf{c}_G(A)) \cdot r_D^{\mathcal{Q}}(A) = \text{ constant } + \sum_{A\subseteq N,\ |A|\geq 2} \mathsf{c}_G(A) \cdot r_D^{\mathcal{Q}}(A) \,.$$

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

Remark. Because adjacencies and immoralities characterize Markov equivalence (see Section 2.2) it follows from Corollary 2 that two acyclic directed graphs over N are Markov equivalent iff the components of their characteristic imsets for sets of cardinality at most three coincide. In particular, the characteristic imset c_G for an acyclic directed graph G over N is uniquely determined by its components for $S \subseteq N$, $2 \leq |S| \leq 3$. Therefore, the restricted characteristic imset (to sets of cardinality at most three) is also a unique BN vector representative of a polynomial length in $|N|$. The reader may think that one can perhaps omit the "superfluous" components of c_G for sets of cardinality four and more. However, the situation is not so easy. The point is that the procedure for computing the remaining components of c_G from those for sets of cardinality at most three is non-linear. More specifically, for a set $S \subseteq N$ of cardinality at least four, one has $\mathsf{c}_G(S) = 1$ iff at least three subsets $T \subset S$ of cardinality $|S| - 1$ exist such that $\mathsf{c}_G(T) = 1$, see Lemma 4.1 [30] for details. Therefore, if we omit the "superfluous" components of c_G the quality criterion $\mathcal{Q}$ becomes a non-linear function of the restricted characteristic imset.

3.2. Inclusion in terms of characteristic imsets

Another interesting question is whether the inclusion of BN structures (see Section 2.2) can be recognized on the basis of the corresponding characteristic imsets. This is indeed the case.

The first step is to characterize, for acyclic directed graphs K and L over N, the *maximal non-trivial inclusion* of the BN structure defined by L in the BN structure defined by K. This is the situation when the statistical model induced by L (= the BN structure determined by L) is strictly contained in the statistical model induced by K while there is no acyclic directed graph G over N such the statistical model induced by G is strictly between the statistical models induced by L and K. This graphically corresponds to the situation when there exists an acyclic directed graph K' Markov equivalent to K and an acyclic directed graph L' equivalent to L such that L' is obtained from K' by the removal of one arrow (Lemma 8.5 in [22] or [5]).

Lemma 2. *Let K, L be two acyclic directed graphs over N. Then the BN structure defined by L is maximally non-trivially included in the BN structure defined by K iff there exists an elementary CI statement $a \perp\!\!\!\perp b \mid C$ such that $\mathsf{c}_K - \mathsf{c}_L$ is the (upper) portrait of $\mathsf{u}_{\langle a,b|C\rangle}$, that is, $\mathsf{c}_K - \mathsf{c}_L = \mathsf{p}_{\langle a,b|C\rangle}$.*

Proof. It follows from Remark 8.10 and Corollary 8.4 in [22] that the above mentioned inclusion relation between statistical models is equivalent to the condition that $\mathsf{u}_L - \mathsf{u}_K$ is an elementary imset $\mathsf{u}_{\langle a,b|C\rangle}$. Now, it remains to apply the invertible

Fig. 1. An example that $c_K \geq c_L$ does not imply the inclusion.

linear portrait mapping defined in (4) and obtain

$$u_{\langle a,b|C\rangle} = u_L - u_K \;\Leftrightarrow\; p_{\langle a,b|C\rangle} = p_L - p_K = (1-p_K) - (1-p_L) \overset{(5)}{=} c_K - c_L,$$

which gives the desired conclusion. □

As the portrait of an elementary imset is a non-negative vector, the consequence of (repeated application of) Lemma 2 is that if the BN structure defined by K includes the one defined by L then $c_K - c_L \geq 0$. This is a simple necessary condition for the inclusion, but not a sufficient one. The counterexample is in Figure 1: we have

$$c_K \equiv \delta_{\{a,b,c\}} + \delta_{\{a,c\}} + \delta_{\{b,c\}} \geq \delta_{\{a,c\}} + \delta_{\{b,c\}} \equiv c_L,$$

but the BN structures defined by K and L are not in the inclusion relation.

Corollary 3. *Let K, L be two acyclic directed graphs over N. Then the BN structure determined by K contains the BN structure determined by L iff $c_K - c_L$ is the combination of portraits of elementary imsets with non-negative integer coefficients.*

Proof. Since there is a finite number of acyclic directed graphs over N, the inclusion premise implies that there exists a sequence of acyclic directed graphs $K = G_1, \ldots, G_n = L, n \geq 1$ such that, for $i = 1, \ldots n-1$, G_i and G_{i+1} are in the relation mentioned in Lemma 2. Conversely, if $c_K - c_L$ is a combination of portraits of elementary imsets, then Lemma 2 implies the inclusion. This is because adding the portrait of an elementary imset to a characteristic imset corresponds to the removal of an arrow. □

In particular, the inclusion of BN structures can be tested by tools of linear programming. Corollary 3 reduces it to testing whether $\{\boldsymbol{x} \geq 0;\ \boldsymbol{A}\boldsymbol{x} = c_K - c_L\} \neq \emptyset$ for a zero-one matrix $\boldsymbol{A}$. Indeed, the columns of $\boldsymbol{A}$ are portraits of elementary imsets.

4. Transition between graphs and characteristic imsets

4.1. From a graph to the characteristic imset

Now, 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 over N. For any $|S| \geq 1$ one has $c_G(S) = 1$ iff*

$$\exists\, \emptyset \neq K \subseteq S \text{ complete in } H, \text{ with } S \setminus K \subseteq pa_H(K). \tag{10}$$

Proof. In an acyclic directed graph G, the only non-empty complete sets are singletons. Thus, by Theorem 1, $c_G(S) = 1$ iff (10) 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 $S \subseteq N$, $|S| \geq 1$, (10) holds with H iff it holds with $\tilde{H}$. To verify this observe that any set S satisfying (10) 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 (10) 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 verify that this change satisfies condition (10) 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 $S \subseteq N$, $|S| \geq 1$, (10) with G is equivalent to (10) with G^*, and the latter is equivalent to (10) with H. □

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

4.2. Back to the essential graph

Corollary 2 allows us to reconstruct the essential graph from the characteristic imset. Indeed, the conditions (i) and (ii) there determine both the adjacencies and immoralities (in any acyclic directed graph G defining the corresponding BN

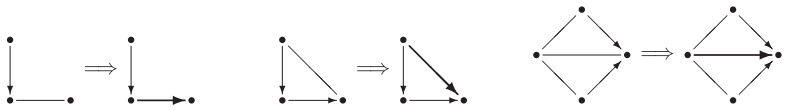

Fig. 2. Orientation rules for getting the essential graph.

structure). Thus, we can directly get the *pattern* (of G) which is the graph having the same adjacencies as G, with arrows belonging to immoralities (in G) directed as in G and the remaining adjacencies being undirected edges.

Thus, the pattern is shared by all acyclic directed graphs defining the same BN structure. Nevertheless, it neither has to be the essential graph nor even a chain graph. However, there is a simple (polynomial-time) procedure for transforming the pattern into the corresponding essential graph G^*. It consists of the (repeated) application of three orientation rules. Specifically, Theorem 3 in [15] states that the exhaustive application of rules from Figure 2 to the pattern of an acyclic directed graph G results in the essential graph G^* (of the equivalence class $\mathcal{G}$ containing G).

5. Learning decomposable models

Note that every chain graph G (over N) can also be interpreted as a statistical model: the class of Markovian probability distributions on the joint sample space $\prod_{i\in N} \mathsf{X}_i$ can be ascribed to G [13]. Analogously, one can extend the concept of Markov equivalence to chain graphs. An undirected graph is Markov equivalent to an acyclic directed graph iff it is chordal [2]. More specifically, a chordal undirected graph G is equivalent to any acyclic directed graph without immoralities obtained by directing edges of G; the essential graph is then G. Therefore, *learning decomposable models* (= chordal undirected graphs) can be viewed as a special case of learning BN structure, specifically, it is learning restricted to a particular subclass of BN structures.

Corollary 4. *Let H be a chordal undirected graph over N. Then the corresponding characteristic imset c_H is specified as follows:*

$$\mathsf{c}_H(A) = 1 \text{ iff } A \text{ is a complete set in } H.$$

Proof. Consider the equivalence class $\mathcal{G}$ of acyclic directed graphs equivalent to H and apply Theorem 2. Since H has no arrows, (10) is equivalent to the above requirement. □

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

Corollary 5. *Let H be an undirected forest over N. Then the corresponding characteristic imset c_H vanishes for sets of cardinality three and more, and, for distinct $a, b \in N$, we have $\mathsf{c}_H(\{a,b\}) = 1$ 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 $\binom{|N|}{2}$, namely the characteristic vector of its edge set. Actually, this motivated our terminology.

The above observation simplifies many things. For example, if maximizing a quality criterion $\mathcal{Q}$ over (undirected) forests is of interest, then, by Lemma 1, the function

$$\mathsf{c}_H \in \mathbb{Z}^{\mathcal{P}_2(N)} \quad\longmapsto\quad \langle r_D^{\mathcal{Q}}, \mathsf{c}_H\rangle = \sum_{A \text{ edge in } H} r_D^{\mathcal{Q}}(A) = \sum_{A \text{ edge in } H} t_D^{\mathcal{Q}}(A)$$

should be maximized, since $r_D^{\mathcal{Q}}(A) = t_D^{\mathcal{Q}}(A)$ for $|A| = 2$ by (8).

In particular, in case of the MLL criterion this means maximizing the sum of weights $\sum_{\{a,b\} \text{ edge}} w_{\{a,b\}}$, where $w_{\{a,b\}} = \mathcal{H}(\hat{P}_{\{a,b\}} | \hat{P}_{\{a\}} \times \hat{P}_{\{b\}})$ is the (empirical) *mutual information* between a and b; see Section 2.4.1.

The polytope spanned by (restricted) characteristic imsets for forests has already been studied in matroid theory [20] and appears to be quite nice from an algorithmic point of view. One important observation is the existence of a simple polynomial-time procedure based on the *greedy algorithm* for finding a 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 [6]. Taking into account what was said above, their procedure can be interpreted as the maximization of the MLL criterion over spanning trees using the greedy technique.

6. Related work

In the literature, we came across two papers in which zero-one vectors (of an exponential length in $|N|$) were used for similar purposes as we (plan to) use characteristic imsets. Moreover, during the reviewing process, two other relevant papers

have been published and the reviewers kindly attracted our attention to them. In this section, we relate our approach to those papers.

6.1. Niepert's LP method for testing CI implications

The paper by Niepert [17] is devoted to testing CI implications by tools of LP. More specifically, given a triplet of pairwise disjoint sets of variables $A, B, C \subseteq N$ the corresponding CI statement $A \perp\!\!\!\perp B \mid C$ is encoded by a zero-one vector as follows. First, Niepert introduces a special class of sets, called the *semi-lattice* of $A \perp\!\!\!\perp B \mid C$:

$$\mathcal{L}(A, B|C) := \{ S \subseteq N;\ \ C \subseteq S \text{ and } A \setminus S \neq \emptyset \neq B \setminus S \},$$

and then defines the vector

$$\mathbf{v}_{\langle A,B|C\rangle}(S) = \begin{cases} 1 & \text{if } S \in \mathcal{L}(A, B|C)\,, \\ 0 & \text{if } S \notin \mathcal{L}(A, B|C)\,, \end{cases} \qquad \text{for } S \subseteq N.$$

It is no problem to see (we leave it to the reader) that $\mathbf{v}_{\langle A,B|C\rangle}$ is, in fact, the *lower portrait* of the semi-elementary imset $\mathrm{u}_{\langle A,B|C\rangle}$:

$$\mathbf{v}_{\langle A,B|C\rangle}(S) = \sum_{T\subseteq S} \mathrm{u}_{\langle A,B|C\rangle}(T) \qquad \text{for any } S \subseteq N.$$

The difference from our upper portrait is that here the sum is over subsets of S, while in Definition 1 we sum over supersets of S. An additional formal difference is that in Definition 1 we ignore components for $|S| \leq 1$ because they automatically vanish, while in case of $\mathbf{v}_{\langle A,B|C\rangle}$ the components for S with $|S| \geq |N| - 1$ always vanish.

Since both portrait mappings are linear and invertible (by well-known Möbius inversion), there exists a one-to-one linear transformation ascribing $\mathrm{p}_{\langle A,B|C\rangle}$ to $\mathbf{v}_{\langle A,B|C\rangle}$. As the characteristic imset for the acyclic directed graph G corresponding to $A \perp\!\!\!\perp B \mid C$ (see Section 2.4.2) is an invertible affine function of $\mathrm{p}_G \equiv \mathrm{p}_{\langle A,B|C\rangle}$, one can obtain the characteristic imset c_G from $\mathbf{v}_{\langle A,B|C\rangle}$ by an invertible affine transformation. Of course, this only concerns acyclic directed graphs encoding CI statements.

The reader may be interested in why we decided to use the upper portrait in Definition 1 and not the lower one. Both transformations are equally suitable if one limits one's attention to CI statements. However, since our aim is learning BN structure we wished to have zero-one vector representatives for all of them. The lower portrait transformation is not suitable from this point of view because it may transform standard imsets outside zero-one vectors. For example, consider the case $N = \{a, b, c\}$ and the empty graph G over N. Then $\mathrm{u}_G = \delta_N - \delta_{\{a\}} - \delta_{\{b\}} - \delta_{\{c\}} + 2 \cdot \delta_\emptyset$ and the corresponding (lower) portrait representative $\mathbf{v}_G$ is

$$\mathbf{v}_G = \delta_{\{a\}} + \delta_{\{b\}} + \delta_{\{c\}} + 2 \cdot \delta_\emptyset\,.$$

6.2. Jaakkola et al.' LP approach to learning BN structure

Like our paper, the paper by Jaakkola et al. [11] was devoted to the application of methods of polyhedral geometry to learning BN structure. They have used the following straightforward zero-one encoding of (acyclic) directed graphs: the components of their vectors are indexed by pairs $(i|B)$, where $i \in N$ and $B \subseteq N \setminus \{i\}$. Given an acyclic directed graph G over N, the respective vector $\boldsymbol{\eta}_G$ is defined as follows:

$$\eta_G(i|B) = \begin{cases} 1 & \text{if } B = pa_G(i),\ i \in N, \\ 0 & \text{otherwise.} \end{cases}$$

The point is that then one can re-write (1) in the form

$$\mathcal{Q}(G, D) \ = \sum_{i\in N} \sum_{B\subseteq N\setminus\{i\}} q_{i|B}(D_{\{i\}\cup B}) \cdot \eta_G(i|B)\,, \tag{11}$$

which allows one to interpret $\mathcal{Q}$ as a linear function of $\boldsymbol{\eta}_G$. Thus, they also transformed the task of maximizing $\mathcal{Q}$ to an LP problem.

Moreover, Jaakkola *et al.* have provided an explicit LP relaxation of their polytope (spanned by vectors $\boldsymbol{\eta}_G$ for acyclic directed graphs). Besides evident non-negativity $\eta(i|B) \geq 0$ and equality constraints $\sum_{B\subseteq N\setminus\{j\}} \eta(j|B) = 1$, for any $j \in N$, they introduced so-called *cluster inequalities*

$$1 \leq \sum_{i\in C} \sum_{B\subseteq N\setminus C} \eta(i|B)\,, \tag{12}$$

which encode acyclicity restrictions on G. The point is that every lattice point in the polyhedron specified by those inequalities is already the vector $\boldsymbol{\eta}_G$ for some acyclic directed graph G over N.

This allows one to use methods of *integer programming* (IP). To avoid the exponential length of $\boldsymbol{\eta}$-vectors in $|N|$, Jaakkola *et al.* used an important pre-processing step, based on observations from [8]. The idea of *pruning* of the components of $\boldsymbol{\eta}$ is based on the observation that in practice one can often exclude from consideration huge parent sets, because of a particular form of the database and the criterion; see Section 6.4 for further details.

To cope with the exponential number of cluster inequalities in $|N|$ they used the cutting plane approach. Thus, they have not applied all the inequalities (12) simultaneously; instead, they add a particular cluster inequality only when it appears to be convenient. What was special in their approach is that they used the dual LP problem formulation as a tool for guiding which additional cluster constraint to add.

Of course, while $\boldsymbol{\eta}_G$ is in a one-to-one correspondence to G, the characteristic imset c_G is not, because it corresponds to the Markov equivalence class of graphs. Thus, there is no mapping transforming c_G to $\boldsymbol{\eta}_G$.

On the other hand, c_G can be viewed as a linear function of $\boldsymbol{\eta}_G$. In [29], which should be a basis of a future paper, the following formula was derived:

$$\mathsf{c}_G(S) = \sum_{i\in S} \sum_{B,\, S\setminus\{i\}\subseteq B\subseteq N\setminus\{i\}} \eta_G(i|B) \qquad \text{for } |S| \geq 1. \tag{13}$$

In general, the transformation of linear inequalities through a many-to-one linear mapping is a complicated mathematical task. Nevertheless, the image of the $\boldsymbol{\eta}$-polyhedron specified by the above inequalities through (13) was characterized in terms of linear inequalities in [29]. Luckily, the cluster inequalities can easily be transformed to the framework of characteristic imsets:

$$1 \leq |C| - \sum_{S\subseteq C,\ |S|\geq 2} \mathsf{c}(S)\cdot(-1)^{|S|}\,. \tag{14}$$

However, paradoxically, the problem occurs with the transformation of basic non-negativity and equality constraints. Because of the many-to-one correspondence $\boldsymbol{\eta}_G \mapsto \mathsf{c}_G$, the number of corresponding inequalities for c is higher than the number of inequalities for $\boldsymbol{\eta}$, which is the price for having unique representatives of BN structures.

An interesting fact is that some of the basic inequalities $\mathsf{c}_G(S) \leq 1$ are not implied by the transformed inequalities for $\boldsymbol{\eta}_G$. Another non-trivial important observation from [29] is that the transformed linear inequalities define an LP relaxation of the characteristic imset polytope.

6.3. Cussens' LP approach to learning BN structure

Cussens [7] also applied the IP approach to structural learning Bayesian networks. He used the same way of vector encoding of (acyclic) directed graphs as Jaakkola et al. [11]. Being inspired by them, he also utilized the cluster inequalities (12) and the cutting plane approach. Of course, he also took the advantage of the idea of pruning the components of the $\boldsymbol{\eta}$-vector.

However, unlike Jaakkola et al., Cussens have not used the dual LP formulation. Instead, he utilized specialized IP-solving software SCIP, which has build-in strategies for finding cutting planes. Besides the inequalities (12) he considered their generalization called *k-cluster-based* constraints:

$$k \leq \sum_{i\in C} \sum_{B\subseteq N\setminus\{i\},\ |B\cap C|<k} \eta(i|B)\,,$$

where $C \subseteq N$ and $1 \leq k \leq |C|$ is an integer, and additional cutting planes called *Gomory cuts*, offered by the integer programming theory.

Cussens also introduced special *surplus variables* s^k_C meaning how far the right-hand side of the above inequality is above its lower bound k. He mentioned an interesting connection of the surplus variables for $2 \leq |C| \leq 3$ and $k = 1$ with adjacencies and immoralities in the respective (acyclic directed) graph G, which similar to our Corollary 2. More specifically, he mentioned

- a and b are adjacent in G iff $s^1_{\{a,b\}} = 0$,
- $a \to c \leftarrow b$ is an immorality in G iff $s^1_{\{a,c\}} = 0$, $s^1_{\{b,c\}} = 0$, $s^1_{\{a,b\}} = 1$ and $s^1_{\{a,b,c\}} = 1$.

One of the reviewer was interested in whether there is a possible connection between the surplus variables and the characteristic imsets. Indeed, there is a relation which easily follows from the cluster inequality (14) in terms of the characteristic imset. If $C = \{a, b\}$ one has, by (14), $s^1_{\{a,b\}} = 1 - \mathsf{c}_G(\{a, b\})$, which means that Cussens' observation about adjacencies is

equivalent to (i) in Corollary 2. In the case $C = \{a, b, c\}$ one has, by (14),

$$s^1_{\{a,b,c\}} = 2 - c_G(\{a, b\}) - c_G(\{a, c\}) - c_G(\{b, c\}) + c_G(\{a, b, c\}) .$$

Thus, Cussens' immorality condition in terms of c_G has the form $c_G(\{a, c\}) = c_G(\{b, c\}) = 1$, $c_G(\{a, b\}) = 0$ and $c_G(\{a, b, c\}) = 1$, which means his observation about immoralities follows from (ii) in Corollary 2.

6.4. The idea of pruning

This is an idea explicated in [8] and then applied in [11,7]. The basic observation, proved as Lemma 1 in [9], which is an extended version of [8], is as follows. If $\mathcal{Q}$ be an additively decomposable criterion and D a database such that, for some $i \in N$ and $B \subseteq N \setminus \{i\}$,

$$\exists\, C \subset B \quad q_{i|C}(D_{\{i\}\cup C}) > q_{i|B}(D_{\{i\}\cup B}) , \tag{15}$$

then there is no acyclic directed graph G over N maximizing $G \mapsto \mathcal{Q}(G, D)$ with $pa_G(i) = B$. The idea of the proof is to consider the graph H obtained from G by the removal of arrows from nodes in $B \setminus C$ to i and observe $\mathcal{Q}(H, D) > \mathcal{Q}(G, D)$.

Thus, provided the condition (15) holds for some i and B, one cannot have an optimal acyclic directed graph G with $pa_G(i) = B$, for which reason, the component of $\boldsymbol{\eta}_G$ for $(i|B)$ and the respective local score $q_{i|B}(D_{\{i\}\cup B})$ can be excluded from the considerations in (11).

However, to prune most of the components of $\boldsymbol{\eta}$ one has to verify exponentially many such conditions (15). The point is that the criteria used in practice somehow prefer sparse graphs. For example, the BIC score has a penalty term which, in fact, protects huge parent sets to occur in the optimal graph. This was the observation made already in Theorem 4.6 of [3], saying that the size of the parent set in an optimal acyclic directed graph with respect to BIC has $\mathcal{O}(\ln \ell)$ upper bound, where ℓ is the length of the database D. Analogously, in [9], stronger sufficient conditions for pruning with BIC were derived which allow, for some $i \in N$ and $B \subseteq N \setminus \{i\}$, to prune with $(i|B)$ also all pais $(i|B')$ where $B \subseteq B' \subseteq N \setminus \{i\}$. Besides the conditions for BIC, also one such a condition for the BDE score was obtained in [9].

Actually, as reported in §6 of [9], the pruning procedure was applied to some databases from so-called UCI repository, and it typically resulted in the reduction of the parent set cardinality to at most 5; only in a few cases the maximal parent set cardinality was 7 or 8.

The reader may be interested in whether pruning can be utilized in the context of characteristic imsets. It follows from what said above and the formula (13) that if, for $S \subseteq N$, $|S| \geq 2$, the condition (15) holds for any $i \in S$ and $B \subseteq N \setminus \{i\}$ with $S \setminus \{i\} \subseteq B$, then $c_G(S) = 0$ for any acyclic directed graph G over N maximizing $G \mapsto \mathcal{Q}(G, D)$. In particular, if $\boldsymbol{\eta}$-vector has been pruned in such a way that there is no component $\eta(i|B)$ in it with $|B| > k$, then one can assume that there is no component $c(S)$ with $|S| > k + 1$. Thus, the result of the pruning procedure can be utilized in our framework, too.

Another relevant question is how to get the components of the (revised) data vector on the basis of (pruned) local scores. Provided $\mathcal{Q}$ be a score equivalent and additively decomposable criterion with local scores $q_{i|B}(*)$, one can use the following procedure (the proof of its correctness is omitted in this paper). First, the local scores are standardized:

$$\hat{q}_{i|B}(D_{\{i\}\cup B}) = q_{i|B}(D_{\{i\}\cup B}) - q_{i|\emptyset}(D_{\{i\}}) \qquad \text{for } i \in N,\ B \subseteq N \setminus \{i\} .$$

The components of the standardized data vector $t^{\mathcal{Q}}_D$ from Section 2.4.1 can be then computed as follows. For $T \subseteq N$, $|T| \leq 1$ put $t^{\mathcal{Q}}_D(T) = 0$, while for $T \subseteq N$, $|T| \geq 2$ consider any total order ρ of the elements in T and introduce:

$$t^{\mathcal{Q}}_D(T) = \sum_{i\in T} \hat{q}_D(i \mid B^\rho_i) , \quad \text{where } B^\rho_i \text{ denotes the set of predecessors of } i \text{ in } \rho.$$

Note that the right-hand expression above does not depend on the choice of ρ because of the assumption $\mathcal{Q}$ is score equivalent. Finally, one can use (8) to compute the components of the revised data vector $r^{\mathcal{Q}}_D(S)$.

7. Preliminary computational experiments

In [14] an indirect LP relaxation via an extension of the characteristic imset polytope has been introduced. For this purpose, any characteristic imset vector c_G was extended with the incidence vector y_G of arrows in an acyclic directed graph G from the Markov equivalence class defined by c_G. That means, extended vectors (y_G, c_G) were considered, where c_G is the characteristic imset for G. The extended encoding has the advantage that the acyclicity of the graph G can easily be ensured by linear restrictions on y_G. The second set of inequalities then implies a correct definition of the characteristic imset vector c_G based on the graph G defined by y_G. In particular, an iterative extension of $c_G(S)$ for sets S with $|S| \geq 4$ is gained; cf. Remark following Lemma 1 and see Lemma 4.1 in [30] or Chapter 2 in [14] for details.

Some other advantageous features of this extended encoding has been described in [14], too. Additional inequalities, more advanced formulations (preserving integrality on some variables) and the descriptions of polytopes for learning restricted

classes of BN structures can easily be obtained. However, the simplification in the description, like pruning (see Section 6.4), to reduce both the number of inequalities and variables, is necessary and, hence, implemented. If pruning is not applicable or does not yield enough short description, column generation still remains to be an option for solving large instance problems.

For some of these LP relaxations, preliminary computational tests for learning BN structure have been performed and described in [14]; the experiments were based on several freely available databases, such as the already mentioned UCI repository, with up to $|N| = 25$ variables. The obtained optimal BN structures have then been analyzed with respect to the used ways of the description and running times. The results of these experiments show that the simplification in the description appears to be particularly necessary and useful.

Using state-of-the-art software like CPLEX, quick optimization procedures, especially fitted to binary problems, can be exploited. The preliminary computational tests validate the applicability of our approach and the strength of the state-of-the-art optimization software for learning BN structure and leave the potential for further improvement in running times.

8. Conclusions

Let us summarize the main contributions of the paper. We came with a new simple unique representative of the BN structure, which opens a combinatorial view on the learning task. It is much closer to the graphical description: there is an easy transition from any graph defining the BN structure to the characteristic imset. We also believe that the procedure for recovering the essential graph from the characteristic imset is much simpler than an analogous procedure in the case of the standard imset described in [24].

The new point of view also gives an elegant geometric interpretation to the classic learning procedure for (spanning) trees from [6]. Of course, this allows one to generalize the greedy learning procedures to other criteria (like the BIC criterion) and to undirected forests. However, this is not the point because this was more or less known in the probabilistic reasoning community.

What seems to be more promising is the potential of future application of advanced methods of linear and integer programming to learning BN structures. To apply the standard LP methods like the *simplex method* one would need to find a polyhedral description of the characteristic imset polytope, which is an open problem. Nevertheless, to apply some advanced methods of *integer programming* it is enough to find a suitable LP relaxation of the polytope, and we already have some LP relaxations [29,14].

Further research topic could be the study of sub-polytopes of the characteristic imset polytope, which may result in LP/IP methods for learning special subclasses of the BN structures. For example, we hope that characteristic imsets can be applied successfully to learning decomposable models with an upper bound on the cardinality of cliques. Because learning decomposable models with cliques of cardinality at most three is already NP-hard (cf. [30]), this can appear to be a non-trivial generalization of the greedy procedure for learning undirected forests.

The idea of a suitable transformation of vector representatives also appears to be useful in the context of testing CI implications by the LP method. In [4], CI statements were encoded by semi-elementary imsets, but as indicated in Section 6.1, additional lower portrait transformation leads to zero-one vector representatives [17]. The use of the upper portrait transformation does result in similar computational speed-ups as in [17], see [14] for details.

Acknowledgements

The research of Milan Studený has been supported by the Grants GAČR n. 201/08/0539 and MŠMT n. 1M0572. We would also like to thank to both reviewers of the paper for their valuable comments and suggestions which helped to improve the quality of the paper.

References

[1] S.A. Andersson, D. Madigan, M.D. Perlman, A characterization of Markov equivalence classes for acyclic digraphs, Annals of Statistics 25 (1997) 505–541.
[2] S.A. Andersson, D. Madigan, M.D. Perlman, On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs, Scandinavian Journal of Statistics 24 (1997) 81–102.
[3] R.R. Bouckaert, Bayesian belief networks: from construction to evidence, Ph.D. thesis, University of Utrecht, 1995.
[4] R.R. Bouckaert, R. Hemmecke, S. Lindner, M. Studený, Efficient algorithms for conditional independence inference, Journal of Machine Learning Research 11 (2010) 3453–3479.
[5] D.M. Chickering, Optimal structure identification with greedy search, Journal of Machine Learning Research 3 (2002) 507–554.
[6] C.K. Chow, C.N. Liu, Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory 14 (1968) 462–467.
[7] J. Cussens, Bayesian network learning with cutting planes, in: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2011, pp. 153–160.
[8] C.P. de Campos, Z. Zeng, Q. Ji, Structure learning Bayesian networks using constraints, in: Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, Canada, 2009, pp. 113–120.
[9] C.P. de Campos, Q. Ji, Efficient structure learning Bayesian networks using constraints, Journal of Machine Learning Research 12 (2011) 663–689.
[10] M. Frydenberg, The chain graph Markov property, Scandinavian Journal of Statistics 17 (1990) 333–353.
[11] T. Jaakkola, D. Sontag, A. Globerson, M. Meila, Learning Bayesian network structure using LP relaxations, in: JMLR Workshop and Conference Proceedings, vol. 9: AISTATS, 2010, pp. 358–365.
[12] D. Heckerman, D. Geiger, D.M. Chickering, Learing Bayesian networks: the combination of knowledge and statistical data, Machine Learning 20 (1995) 194–243.

[13] S.L. Lauritzen, Graphical Models, Clarendon Press, 1996.
[14] S. Lindner, Discrete optimization in machine learning – learning Bayesian network structures and conditional independence implication, Ph.D. thesis, TU Munich, 2012.
[15] C. Meek, Causal inference and causal explanation with background knowledge, in: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 1995, pp. 403–410.
[16] R.E. Neapolitan, Learning Bayesian Networks, Pearson Prentice Hall, 2004.
[17] M. Niepert, Logical inference algorithms and matrix representations for probabilistic conditional independence, in: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 428–435.
[18] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, 1988.
[19] A. Schrijver, Theory of Linear and Integer Programming, John Wiley, 1986.
[20] A. Schrijver, Combinatorial Optimization – Polyhedra and Efficiency, vol. B: Springer Verlag, 2003.
[21] G.E. Schwarz, Estimation of the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[22] M. Studený, Probabilistic Conditional Independence Structures, Springer Verlag, 2005.
[23] M. Studený, Mathematical aspects of learning Bayesian networks: Bayesian quality criteria, research report no. 2234, Institute of Information Theory and Automation, Prague, December 2008.
[24] M. Studený, J. Vomlel, A reconstruction algorithm for the essential graph, International Journal of Approximate Reasoning 50 (2009) 385–413.
[25] M. Studený, A. Roverato, Š. Štěpánová, Two operations of merging and splitting components in a chain graph, Kybernetika 45 (2009) 208–248.
[26] M. Studený, J. Vomlel, R. Hemmecke, A geometric view on learning Bayesian network structures, International Journal of Approximate Reasoning 51 (2010) 578–586.
[27] M. Studený, R. Hemmecke, S. Lindner, Characteristic imset: a simple algebraic representative of a Bayesian network structure, in: P. Myllymäki, T. Roos, T. Jaakkola (Eds.), Proceedings of the 5th European Workshop on Probabilistic Graphical Models, HIIT Publications, 2010, pp. 257–264.
[28] M. Studený, J. Vomlel, On open questions in the geometric approach to structural learning Bayesian nets, International Journal of Approximate Reasoning 52 (2011) 627–640.
[29] M. Studený, D. Haws, On polyhedral approximations of polytopes for learning Bayes nets, research report no. 2303 Institute of Information Theory and Automation of the ASCR, Prague, July 2011. Also available on <http://arxiv.org/abs/1107.4708>.
[30] M. Studený, D. Haws, R. Hemmecke, S. Lindner, Polyhedral approach to statistical learning graphical models, in: Proceedings of the 2nd CREST–SBM International Conference Harmony of Gröbner Bases and the Modern Industrial Society, World Scientific, 2012, pp. 346–372, in press.
[31] T. Verma, J. Pearl, Equivalence and synthesis of causal models, in: Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence, Elsevier, 1991, pp. 220–227.