Integer Linear Programming Approach to Learning Bayesian Network Structure: towards the Essential Graph

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

Abstract

The basic idea of a geometric approach to learning a Bayesian network (BN) structure is to represent every BN structure by a certain vector. If the vector representative is chosen properly, it allows one to re-formulate the task of finding the global maximum of a score over BN structures as an integer linear programming (ILP) problem. Suitable such a zero-one vector representative is the *characteristic imset*, introduced in (Studený, Hemmecke and Lindner, 2010). In this paper, extensions of characteristic imsets are considered which additionally encode chain graphs without flags equivalent to acyclic directed graphs. The main contribution is the polyhedral description (= in terms of a set of linear inequalities) of the respective domain of the ILP problem. It is just a theoretical result, but it opens the way to the application of ILP software packages in the area of learning a BN structure. The advantage of this approach is that, as a by-product of the ILP optimization procedure, one may get the *essential graph*, which is a traditional graphical BN representative.

1 Introduction

Learning *Bayesian network* (BN) structure by a score and search method means to maximize a *quality criterion* $\mathcal{Q}$, also named a *score*, which is a real function of the (acyclic directed) graph G and the observed database D. The value $\mathcal{Q}(G, D)$ evaluates how the BN structure defined by the graph G fits the database D.

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

The geometric approach is to represent every BN structure by a certain vector so that such a criterion $\mathcal{Q}$ becomes an affine function of the vector representative. This idea was introduced already by Studený (2005) and then deepened in (Studený, Vomlel and Hemmecke, 2010). A suitable (uniquely determined) such a zero-one vector BN representative seems to be the *characteristic imset*, introduced at last PGM (Studený, Hemmecke and Lindner, 2010).

Jaakkola et al. (2010) and Cussens (2010; 2011) came independently with an analogous geometric approach. The main difference is that they used certain special zero-one vector codes of (acyclic) directed graphs to represent (non-uniquely) BN structures. On the other hand, they made more progress with the practical use of *integer linear programming* (ILP) tools. To overcome technical problems with the exponential length of their vectors they utilized the idea of reduction of the search space from (de Campos et al., 2009), based on a particular form of databases and criteria occurring in practice.

In (Studený and Haws, 2012), both methods of BN structure vector representation were compared and it was found that the characteristic imset can be viewed as a (many-to-one) linear function of the above mentioned codes of directed graphs. Finally, Lindner (2012) performed some preliminary computational experiments based on the characteristic imset ap-

proach; an overview of this approach has been given in (Studený et al., 2012) and (Hemmecke et al., 2012).

In this paper, an extended vector BN representative is introduced, which encodes both the characteristic imset and a certain special graph (equivalent to an acyclic directed graph). The main result is a *polyhedral characterization* of the domain of the respective ILP problem. More specifically, a set of linear inequalities is presented such that the only vectors with integer components in the polyhedron specified by those inequalities are the above mentioned extended vector representatives.

The inequalities are classified in four groups. The number of inequalities in the first two groups is polynomial in the number of variables (= nodes of the graph), while the number of remaining inequalities is exponential. However, provided the length of the vector representatives is limited/reduced to a polynomial number by the idea of from (de Campos et al., 2009), the number of inequalities in the third group can be reduced to a polynomial number as well. The fourth group of inequalities correspond to acyclicity restrictions. In general, they cannot be reduced to a polynomial number, but because of their natural graphical interpretation, the (modified) cutting plane approach may be applied to solve the respective ILP problems.

Another advantage of this extended vector representative is that one can get, as a result of solving the ILP problem, the *essential graph*, which is known as a standard unique graphical BN representative (Andersson, Madigan and Perlman, 1997).

2 Basic concepts

Let N be a finite non-empty set of *variables*; let's assume $|N| \geq 2$ to avoid the trivial case. 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 in this paper have N as the set of nodes and two types of edges between (distinct) nodes $i, j \in N$, namely directed edges, called *arrows*, and denoted like $i \to j$ (or $j \leftarrow i$), and undirected edges, called *lines*, and denoted like $i - j$ (or $j - i$). No multiple edges are allowed between two nodes. If there is an edge between nodes i and j, we say they are *adjacent*. A graph is *undirected* if it only has lines; it is *directed* if it only has arrows.

A *cycle* of the length $m \geq 3$ in a graph H is a sequence of nodes $\rho : i_0, i_1, \ldots, i_m = i_0$, where $i_1, \ldots, i_m$ are distinct nodes, and i_r and i_{r+1} are adjacent in H (for each $r = 0, \ldots, m-1$). The cycle ρ is *chordless* if there is no other edge in H between nodes in $\{i_1, \ldots, i_m = i_0\}$ besides those which form the cycle ρ. An undirected graph is called *chordal* if it has no chordless cycle of the length $m \geq 4$.

The cycle ρ is *directed* if $i_r \to i_{r+1}$ in H for each $r = 0, \ldots, m-1$. A directed graph is *acyclic* if it has no directed cycle (of arbitrary length $m \geq 3$). An equivalent definition of an acyclic directed graph G is that there exists an ordering $b_1, \ldots, b_m$, $m \geq 1$ of all nodes in N which is consistent with the direction of arrows: $b_i \to b_j$ in G implies $i < j$. The set of *parents* of $i \in N$ in a (directed) graph G is the set $pa_G(i) \equiv \{\, j \in N;\ j \to i \text{ in } G\}$.

The cycle ρ is *semi-directed* if $i_0 \to i_1$ in H and, for each $r = 1, \ldots, m-1$ one has either $i_r \to i_{r+1}$ in H or $i_r - i_{r+1}$ in H. A *chain graph* is a graph without semi-directed cycles.

A set $C \subseteq N$ is *connected* if every pair of distinct nodes in C is connected via an undirected path. Maximal connected sets are called *components*. An equivalent definition of a chain graph H is that there exists an ordering $C_1, \ldots, C_m$, $m \geq 1$ of all its components such that if $a \to b$ in H then $a \in C_i$ and $b \in C_j$ with $i < j$.

2.2 Bayesian network structures

In statistical context, each variable $i \in N$ is assigned a finite *sample space* X_i (of possible values); assume $|\mathsf{X}_i| \geq 2$ for each $i \in N$ to avoid technical problems.

A *Bayesian network* (BN) is a pair (G, P), where G is an acyclic directed graph with the node set N and P a *Markovian* probability distribution (with respect to G) on the *joint sample space* $\prod_{i \in N} \mathsf{X}_i$. This means P satisfies condi-

tional independence restrictions determined by G; see (Lauritzen, 1996) for details. The *BN structure* defined by an acyclic directed graph G is the class of Markovian probability distributions with respect to G on (fixed) $\prod_{i\in N} \mathsf{X}_i$.

However, different graph over N can be *Markov equivalent*, which means they define the same BN structure. Classic graphical characterization of (Markov) equivalent graphs by Verma and Pearl (1991) says they are equivalent iff they have the same adjacencies and the same immoralities. Here, an *immorality* in a graph 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).

Learning BN structure means to determine it on the basis of an observed *database* D, which is a sequence $x_1, \ldots, x_\ell$ of elements of $\prod_{i\in N} \mathsf{X}_i$ ($\ell \geq 1$ is the length of the database). This is often done by maximizing some *quality criterion* (= score), which is a real function $\mathcal{Q}$ of two variables: of an acyclic directed graph G and of a database D. The value $\mathcal{Q}(G, D)$ quantitatively evaluates how well the BN structure defined by the graph G explains the occurrence of the database D.

Because the aim is to learn a BN structure, a natural requirement is that $\mathcal{Q}$ should be *score equivalent*, which means that, given any D,

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

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

Additively *decomposable* criterion is a criterion $\mathcal{Q}$ which can be written as follows:

$$\mathcal{Q}(G, D) = \sum_{i\in N} q_{i|\mathrm{pa}_G(i)}(D_{\{i\}\cup \mathrm{pa}_G(i)}), \qquad (1)$$

where $q_{i|B}$ for $i \in N$, $B \subseteq N \setminus \{i\}$ are some real functions and D_A for $\emptyset \neq A \subseteq N$ denotes the projection of the database D to $\prod_{i\in A} \mathsf{X}_i$. The terms $q_{i|B}(D_{\{i\}\cup B})$ are named *local scores*.

Well-known quality criteria used in practice are Schwarz's (1978) *Bayesian information criterion* (BIC) and the *Bayesian Dirichlet Equivalence* (BDE) score (Heckerman et al., 1995).

2.3 Essential graphs

A kind of standard (unique) graphical representative of a BN structure is the so-called essential graph.

Definition 1. Let $\mathcal{G}$ be a Markov equivalence class of acyclic directed graphs over N. The *essential graph* G^* of $\mathcal{G}$ is defined as follows:

- $a \to b$ in G^* if $a \to b$ in every G from $\mathcal{G}$,
- $a - b$ in G^* if there are graphs G_1 and G_2 in $\mathcal{G}$ with $a \to b$ in G_1 and $a \leftarrow b$ in G_2.

This terminology and the first graphical characterization of essential graphs was given by Andersson, Madigan and Perlman (1997). It implies that every essential graph is a chain graph and has no *flag*, by which is meant an induced subgraph for three nodes $\{a, b, c\}$ in which $a \to b - c$ (and a and c are not adjacent). One can introduce a graphical concept of equivalence for these graphs, which generalizes Markov equivalence of acyclic directed graphs.

Definition 2. Two chain graphs without flags G and H over N are *equivalent* if they have the same adjacencies and immoralities. Given two such graphs, we say that H is *larger* than G if, for any $i, j \in N$, $i \to j$ in H implies $i \to j$ in G.

The following characterization of the essential graphs, proved as Corollary 4 in (Studený, 2004), will be utilized later.

Lemma 1. *Let $\mathcal{G}$ be an equivalence class of acyclic directed graphs over N and $\mathcal{H}$ an equivalence class of chain graphs without flags such that $\mathcal{G} \subseteq \mathcal{H}$. Then G^* is the largest graph in $\mathcal{H}$.*

3 Characteristic imset

This algebraic representative of a BN structure was introduced in (Studený, Hemmecke and Lindner, 2010). For our purpose, the following equivalent definition is suitable.

Definition 3. Let G be an acyclic directed graph over N. The *characteristic imset* for G can be introduced as a zero-one vector c_G with components $\mathsf{c}_G(S)$ where $S \subseteq N$, $|S| \geq 2$, such that $\mathsf{c}_G(S) = 1$ iff

$$\exists\, i \in S \text{ such that } S \setminus \{i\} \subseteq pa_G(i). \qquad (2)$$

The point is that two acyclic directed graphs G and H over N are Markov equivalent if and only if $\mathsf{c}_G = \mathsf{c}_H$; see § 3 of (Hemmecke et al., 2012). Moreover, Corollary 2 in (Hemmecke et al., 2012) implies that, for different $i, j, k \in N$,

(i) i and j are adjacent in G iff $\mathsf{c}_G(\{i,j\}) = 1$,

(ii) $i \to k \leftarrow j$ is an immorality in G iff $\mathsf{c}_G(ijk) = 1$ and $\mathsf{c}_G(ij) = 0$.

In particular, one can observe that the characteristic imset c_G is uniquely determined by its values $\mathsf{c}_G(S)$ for $S \subseteq N$, $2 \leq |S| \leq 3$.

It appears to be suitable to have a formula for the characteristic imset on the basis of any graph H in the class $\mathcal{H}$ from Lemma 1. For this purpose one needs the next auxiliary concept.

Definition 4. We say that a graph H over $S \subseteq N$ has a *super-terminal component* if there exists a non-empty set $K \subseteq S$ such that

- K is a *complete set* in H, which means, for each pair of distinct nodes $i, k \in K$, one has $i - k$ in H,
- $\forall\, j \in S \setminus K \;\; \forall\, i \in K$ one has $j \to i$ in H.

It makes no problem see that a super-terminal component K, if exists, is uniquely determined.

The following result follows directly from Theorem 2 in (Hemmecke et al., 2012):

Lemma 2. *Let H be a chain graph without flags equivalent to an acyclic directed graph G. For any $S \subseteq N$, $|S| \geq 2$ one has $\mathsf{c}_G(S) = 1$ iff the induced subgraph of H for S (denoted by H_S) has a super-terminal component.*

3.1 Straightforward codes of graphs

Jaakkola et al. (2010) and Cussens (2010; 2011) used a special method for vector encoding (acyclic) directed graphs over N. The vector η_G encoding G has components indexed by pairs $(i|B)$, where $i \in N$ and $B \subseteq N \setminus \{i\}$. Specifically, it is defined as follows:

$$\eta_G(i|B) = \begin{cases} 1 & B = pa_G(i), \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

In (Studený and Haws, 2012), the relation of the characteristic imset to this straightforward code of G was established. Actually, c_G is a linear function of η_G given by

$$\mathsf{c}_G(S) = \sum_{i \in S} \sum_{B,\, S\setminus\{i\} \subseteq B \subseteq N\setminus\{i\}} \eta_G(i|B) \qquad (4)$$

for $S \subseteq N$, $|S| \geq 2$. Indeed, (4) follows directly from Definition 3: clearly, at most one node $i \in S$ with $S \setminus \{i\} \subseteq pa_G(i)$ exists in an acyclic directed graph G_S.

4 Integer linear programming

The task to maximize a criterion can be reformulated as an *integer linear programming* (ILP) problem. Indeed, by (1) and (3), every decomposable criterion $\mathcal{Q}$ can be interpreted as a linear function of η_G, where G falls within the class of acyclic directed graphs over N.

Jaakkola et al. (2010) gave a finite list of valid inequalities for η_G's, which characterize them in the sense that the only vectors with integer components satisfying those inequalities are the codes of acyclic directed graphs. Such a domain description, in terms of polyhedral geometry named an *LP relaxation* (of the respective polytope), allows one to turn the learning task into an ILP problem: to optimize a linear function over vectors with integer components within a polyhedron.

Specifically, besides basic non-negativity and equality constraints, Jaakkola et al. (2010) came with the following *cluster inequalities*:

$$1 \leq \sum_{i \in S} \sum_{B \subseteq N\setminus S} \eta_G(i|B) \qquad (5)$$

for any $S \subseteq N$, $|S| \geq 2$. The meaning of the inequality (5) is that the induced subgraph G_S has at least one *initial node*, that is, a node $i \in S$ with no $j \in S$ with $j \to i$ in G. As G_S is acyclic, the existence of such a node is obvious.

To overcome the technical problem with the exponential length (in $|N|$) of vectors η_G the idea of *pruning* of their components was applied. The idea taken from (de Campos et al., 2009) is that a particular form of scoring criteria used in practice allows one to conclude (on the

basis of an observed database) that the optimal graph G has no node $i \in N$ with large $|pa_G(i)|$. Therefore, one can exclude from consideration the respective components of the η-vector. This pruning procedure is time demanding, but useful: as reported in § 6 of (de Campos and Ji, 2011), in practical cases it typically results in the reduction of the parent set cardinality to at most 5, only in a few cases the maximal cardinality was 7 or 8.

To overcome the problem with the exponential number of cluster inequalities Jaakkola et al. (2010) used the method of iterative constraint adding, where they employed the dual formulation (of their approximate LP problems) to guide the choice of a newly added cluster constraint.

Cussens was in (2010) interested in pedigree learning, in which case the parent set cardinality is bounded by 2. However, to ensure the acyclicity of the graph G he used another trick: the idea of *extending* the vector BN representatives. He added some additional components to the (reduced) η_G-vector which allowed him to encode the total order of nodes consistent with the direction of arrows in the graph G. Then he introduced easily an LP relaxation for these extended vector representatives. Actually, the number of the added components and the number of inequalities ensuring acyclicity were both polynomial in $|N|$.

The other paper by Cussens (2011) was inspired by Jaakkola et al. (2010). Unrestricted BN structure learning was the goal and to overcome the problem with the exponential number of these inequalities Cussens used the *cutting plane* approach.

4.1 ILP with characteristic imsets

Lemma 1 in (Hemmecke et al., 2012) says that every score equivalent and additively decomposable criterion $\mathcal{Q}$ has the form

$$\mathcal{Q}(G,D) \;=\; \mathcal{Q}(G^{\emptyset},D) \\ +\sum_{S\subseteq N,\,|S|\geq 2} r_D^{\mathcal{Q}}(S)\cdot \mathsf{c}_G(S)\,, \qquad (6)$$

where $G^{\emptyset}$ is the empty graph over N (= without adjacencies) and $r_D^{\mathcal{Q}}$ uniquely determined vector, depending on the database D only, called the *revised data vector* (relative to $\mathcal{Q}$). Given $\mathcal{Q}$, one usually can derive a mathematical formula for the components of $r_D^{\mathcal{Q}}$. An alternative way, described in (Studený, 2012), is to compute $r_D^{\mathcal{Q}}(S)$ for $S \subseteq N$, $|S| \geq 2$ from local scores.

That means, the criterion $\mathcal{Q}$ can be interpreted as an affine function (= a linear function plus a constant) of the characteristic imset c_G, which is a unique BN representative. Thus, to employ the methods of ILP, one has to come with an LP relaxation for characteristic imsets.

In (Studený and Haws, 2012), we transformed the above-mentioned LP relaxation by Jaakkola et al. (2010) through (4) into the framework of characteristic imsets. A pleasant finding was that the cluster inequality (5) takes a neat form

$$\sum_{T\subseteq S,\,|T|\geq 2} \mathsf{c}(T)\cdot(-1)^{|T|} \leq |S|-1\,. \qquad (7)$$

Another non-trivial observation was that the transformed linear inequalities define an LP relaxation of the characteristic imset polytope. However, since the number of resulting inequalities is super-exponential in $|N|$, which is an unpleasant consequence of the many-to-one transformation, this LP relaxation does not seem to be suitable for practical purposes.

Another important fact is that pruning can also be utilized in the context of characteristic imsets. This follows easily from (4): if the components of η_G were pruned to the maximal parent set cardinality k, then one can assume $\mathsf{c}_G(S) = 0$ for $S \subseteq N$ with $|S| > k+1$ in any optimal graph G. Then $r_D^{\mathcal{Q}}(S)$ for such S need not be computed.

Lindner (2012) came with another LP relaxation of the characteristic imset polytope. She also used the idea of extending BN vector representatives: the components added to the characteristic imset allowed her to encode acyclic directed graphs defining the characteristic imset. Finally, she performed some preliminary computational experiments based on this approach.

5 LP relaxation

In this section we present another LP relaxation for the characteristic imset polytope, also based on the idea of adding components.

5.1 Extended vector representative

The result from Lemma 1 motivated the idea to broaden the class of (non-unique) graphical BN representatives to the class of *chain graphs without flags that are equivalent to acyclic directed acyclic graphs*. Lemma 2 then motivated the following definition.

Definition 5. Let H be a chain graph over N without flags (equivalent to an acyclic directed graph). We ascribe to H a zero-one vector $(\mathsf{a}_H, \mathsf{c}_H)$ with components given as follows:

$$\mathsf{a}_H(i \to j) = 1 \iff i \to j \text{ in } H,$$

where $i, j \in N$ are distinct, and,

$$\mathsf{c}_H(S) = 1 \iff H_S \text{ has a super-terminal component,}$$

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

Thus, the a-part of the vector encodes the presence of arrows $i \to j$ in the graph H (= codes of *arrowheads*), while the c-part is, by Lemma 2, the respective characteristic imset. Note that the number of added components $|N| \cdot (|N| - 1)$ is polynomial in $|N|$.

5.2 The list of inequalities

The inequalities are classified in four groups and none of them is superfluous.

The *basic non-negativity inequalities* are:

(b.1) $\forall\, i, j \in N$ distinct $\quad 0 \leq \mathsf{a}(i \to j)$,

(b.2) $\forall\, S \subseteq N$, $|S| = 3, 4$ $\quad 0 \leq \mathsf{c}(S)$.

The *consistency inequalities* mainly relate the a-part to the c-part: for distinct $i, j, k \in N$

(c.1) $\mathsf{a}(i \to j) + \mathsf{a}(j \to i) \leq \mathsf{c}(ij)$,

(c.2) $\mathsf{c}(ij) \leq 1$,

(c.3) $2 \cdot \mathsf{c}(ijk) \leq 2 \cdot \mathsf{c}(ij) + \mathsf{a}(i \to k) + \mathsf{a}(j \to k)$,

(c.4) $\mathsf{a}(i \to j) + \mathsf{c}(jk) \leq 1 + \mathsf{c}(ijk) + \mathsf{a}(j \to k)$,

(c.5) $\mathsf{a}(i \to j) + \mathsf{c}(jk) + \mathsf{c}(ik) \leq 2 + \mathsf{a}(i \to k) + \mathsf{a}(k \to j)$.

The *extension inequalities* ensure that the c-part is determined by values $\mathsf{c}(S)$, $2 \leq |S| \leq 3$:

(e.1) $\forall\, S \subseteq N$, $|S| \geq 3$
$\sum_{i \in S} \mathsf{c}(S \setminus \{i\}) \leq 2 + (|S| - 2) \cdot \mathsf{c}(S)$,

(e.2) $\forall\, S \subseteq N$, $|S| \geq 4$
$(|S| - 1) \cdot \mathsf{c}(S) \leq \sum_{i \in S} \mathsf{c}(S \setminus \{i\})$.

Finally, the *acyclicity inequalities* only concern the c-part and ensure that the solution is the graph in the considered class:

(a.1) $\forall\, S \subseteq N$, $|S| \geq 4$
$\sum_{T \subseteq S, |T| \geq 2} \mathsf{c}(T) \cdot (-1)^{|T|} \leq (|S| - 1)$.

Observe these are just the transformed cluster inequalities (7). Here is the main result.

Theorem 1. *Let H be a chain graph without flags equivalent to an acyclic directed graph over N. Then the inequalities (b.1)-(b.2), (c.1)-(c.5), (e.1)-(e.2) and (a.1) are valid for the vector $(\mathsf{a}_H, \mathsf{c}_H)$ from Definition 5. Conversely, if a vector (a, c) with integer components satisfies those inequalities, then such (uniquely determined) graph H exists with $(\mathsf{a}, \mathsf{c}) = (\mathsf{a}_H, \mathsf{c}_H)$.*

A complete proof is beyond the scope of a conference paper and can be found in (Studený, 2012); in fact, a stronger result is derived there saying that, if (a.1) is omitted, one still gets a code of a certain graph H, but with possible semi-directed cycles. Theorem 1 also holds with (a.1) replaced by a simplified version, perhaps easier to implement:

(a.1*) $\forall\, S \subseteq N$, $|S| \geq 4$
$\sum_{T \subseteq S, 2 \leq |T| \leq 3} \mathsf{c}(T) \cdot (-1)^{|T|} \leq (|S| - 1)$.

The resulting polyhedron is, however, different. Lindner (2012) mentioned (a.1*), too.

5.3 Interpretation of inequalities

One can give graphical *interpretation* to (most of) the inequalities, on which the proof of their necessity is based.

(c.1) means that if $i \to j$ or $j \to i$ is encoded in a then $[i, j]$ is encoded as an edge in c,

(c.2) means, together with (c.1), that one cannot have simultaneously $i \to j$ and $j \to i$ in H,

(c.3) allows one to conclude that if $\mathsf{c}(ijk) = 1$ then H_{ijk} has a super-terminal component,

(c.4) prevents H to have a flag $i \to j - k$,

(c.5) says that H has not a semi-directed 3-cycle of the form i, j, k, i with $i \to j$,

(e.1) means: if S has at least 3 subsets T with $|T| = |S| - 1$ with $\mathsf{c}(T) = 1$ then $\mathsf{c}(S) = 1$,

(e.2) if $\mathsf{c}(S) = 1$ then at least $|S| - 1$ sets $T \subset S$ with $|T| = |S| - 1$ and $\mathsf{c}(T) = 1$ exists.

(a.1) means, loosely said, that the graph H_S has at least one initial node; it forbids the existence of a chordal semi-directed/undirected cycle composed just of the nodes of S.

5.4 The idea of the sufficiency proof

First, one observes that the above inequalities, together with the assumption that (a, c) has integer components, imply that $0 \leq \mathsf{c}(S) \leq 1$ for any $S \subseteq N$, $|S| \geq 2$. Given such a vector (a, c), define a graph H:

$$\begin{aligned} i \to j \ \text{ in } H \ &\Leftrightarrow \ \mathsf{a}(i \to j) = 1, \\ i - j \ \text{ in } H \ &\Leftrightarrow \ \mathsf{c}(ij) = 1 \ \& \\ &\qquad \mathsf{a}(i \to j) = 0 \ \& \ \mathsf{a}(j \to j) = 0. \end{aligned}$$

Then the inequalities allow one to show that H is 3-acyclic (= has no semi-directed cycle of the length 3) and has no flags. The next step is to show that, for distinct $i, j, k \in N$, $\mathsf{c}(ijk) = 1$ iff H_{ijk} has a super-terminal component. Using (e.1)-(e.2), this observation is extended to any $S \subseteq N$, $|S| \geq 2$ in place of ijk. Finally, (a.1) is used to show that H has no semi-directed or undirected chordless cycle of the length $m \geq 4$.

6 Summary of the whole procedure

A pre-processing step should be the *pruning* in the context of characteristic imsets; see § 4.1. The result should be a cache of values $r_D^{\mathcal{Q}}(S)$ for $S \in \mathcal{T}$, where $\mathcal{T} \subseteq \{S \subseteq N;\ |S| \geq 2\}$ is a class of sets closed under subsets such that $\mathsf{c}_G(S) = 0$ for $S \notin \mathcal{T}$ and any optimal G. The hope is that $\mathcal{T}$ will consist of sets of small cardinality.

The first ILP problem to be solved is to maximize the function

$$(\mathsf{a}, \mathsf{c}) \mapsto \sum_{S \in \mathcal{T}} r_D^{\mathcal{Q}}(S) \cdot \mathsf{c}(S)$$

over the domain of vectors with integral components specified by the inequalities from § 5.2. The number of consistency inequalities is polynomial in $|N|$ and, provided $|\mathcal{T}|$ is small, a small number of extension inequalities is applicable.

The number of acyclicity inequalities cannot be reduced to a polynomial amount, but one can apply the idea of *iterative constraint adding*. In the first iteration, the acyclicity inequalities are omitted. The solution of the respective ILP problem corresponds to a graph H with possible semi-directed/undirected chordless cycles. One can identify minimal sets S of nodes in H forming those cycles. The corresponding acyclicity inequalities are incorporated in the current list of inequalities and a revised ILP problem is solved. This procedure is repeated until either a solution without those cycles is found or memory overfill.

Once such a graph H is found, the respective *essential graph* G^* can be obtained as follows. The idea is to fix the c-part of the vector solution and formulate the second ILP problem: to minimize the objective

$$(\mathsf{a}, \mathsf{c}) \mapsto \sum_{i,j \in N, i \neq j} \mathsf{a}(i \to j)$$

instead, under non-negativity and consistency constraints. By Lemma 1, the solution should correspond to the essential graph G^* for c.

Conclusions

The main theoretical advantage of the presented approach (in comparison with the other ILP approaches) is that the solution is a unique BN representative and the vector representatives are more compressed. The augmentation does not kill this comparative advantage because the extended characteristic imsets are still

more than $\frac{|N|}{3}$-times shorter than the straightforward graph codes. The number of inequalities is also smaller than Lindner (2012) used.

Another fine feature is that the presented approach allows one to get directly the *essential graph* as a result of solving the ILP problem, that is, to avoid in this way the need for additional graph-reconstruction procedure.

Nevertheless, the theoretical assumptions (or ambitions) must be tested in practice. Realize that there is also strong influence of the pruning on the efficiency of the overall procedure and it is not clear at present whether a smaller number of inequalities is better than a tighter LP relaxation involving more inequalities. Thus, the proposed LP relaxation should become a basis for computational experiments, made in cooperation with foreign colleagues.

Acknowledgments

This research has been supported by the grant GAČR n. 201/08/0539.

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.

James Cussens. 2010. Maximum likelihood pedigree reconstruction using integer programming. In *Workshop on Constraint Based Methods for Bioinformatics*, pages 9–19.

James Cussens. 2011. Bayesian network learning with cutting planes. In *27th Conference on Uncertainty in Artificial Intelligence*, pages 153–160.

Cassio P. de Campos, Zhi Zeng and Qiang Ji. 2009. Structure learning Bayesian networks using constraints. In *26th International Conference on Machine Learning*, pages 113–120.

Cassio P. de Campos and Qiang Ji. 2011. Efficient structure learning Bayesian networks using constraints. *Journal of Machine Learning Research*, 12:663–689.

David Heckerman, Dan Geiger and David M. Chickering. 1995. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20:194–243.

Raymond Hemmecke, Silvia Lindner and Milan Studený. 2012. Characteristic imsets for learning Bayesian network structure. To appear in *International Journal of Approximate Reasoning*, see `doi:10.1016/j.ijar.2012.04.001`.

Tommi Jaakkola, David Sontag, Amir Globerson and Marina Meila. 2010. Learning Bayesian network structure using LP relaxations. In *JMLR Workshop and Conference Proceedings, volume 9: AISTATS*, pages 358–365.

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

Silvia Lindner. 2012. Discrete optimization in machine learning - learning Bayesian network structures and conditional independence implication. PhD thesis, TU Munich.

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

Milan Studený, Raymond Hemmecke and Silvia Lindner. 2010. Characteristic imset: a simple algebraic representative of a Bayesian network structure. In *5th European Workshop on Probabilistic Graphical Models*, pages 257–264.

Milan Studený and David Haws. 2012. On polyhedral approximations of polytopes for learning Bayesian networks. Submitted to *Journal of Algebraic Statistics*; previous working paper available on `http://arxiv.org/abs/1107.4708`.

Milan Studený, David Haws, Raymond Hemmecke and Silvia Lindner. 2012. Polyhedral approach to statistical learning graphical models. In *2nd CREST–SBM International Conference*, World Scientific, pages 346–372.

Milan Studený. 2012. LP relaxations and pruning for characteristic imsets. Research report n. 2323, Institute of Information Theory and Automation of the ASCR; available through `staff.utia.cas.cz/studeny/f18.html`.

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