POLYHEDRAL APPROACH TO STATISTICAL LEARNING GRAPHICAL MODELS

Milan Studený

Institute of Information Theory and Automation of the ASCR,
(ASCR = the Academy of Sciences of the Czech Republic)
Pod Vodárenskou věží 4, 18208 Prague, Czech Republic
E-mail: studeny@utia.cas.cz

David Haws

Department of Statistics, University of Kentucky
871 Patterson Office Tower, Lexington, KY 40506-0027, U.S.A.
E-mail: dchaws@gmail.com

Raymond Hemmecke* and Silvia Lindner†

Zentrum Mathematik, Technische Universität Munich
Boltzmannstrasse 3, 85747 Garching, Germany
**E-mail: hemmecke@ma.tum.de*
†*E-mail: slindner@ma.tum.de*

The statistical task to learn graphical models of Bayesian network structure from data leads to the study of special polyhedra. In the paper, we offer an overview of our polyhedral approach to learning these statistical models. First, we report on the results on this topic from our recent papers. The second part of the paper brings some specific additional results inspired by this approach.

Keywords: Bayesian network structure; standard imset; characteristic imset; polyhedral geometry.

1. Introduction

Bayesian networks are popular graphical models, used widely both in statistics and artificial intelligence; see the books by Lauritzen[1] and Pearl[2]. These statistical models of conditional independence structure are ascribed to acyclic directed graphs whose nodes correspond to (random) variables in consideration. An important topic is learning *Bayesian network structure* (BN structure), which is determining the (most suitable) statistical model on the basis of observed data; see the book by Neapolitan[3] for details.

Although there are learning methods based on statistical conditional independence tests, contemporary methods are based on maximization of a suitable *quality criterion*, also named *score* function. It is a function $\mathcal{Q}(G, D)$ of the (acyclic directed) graph G and the data D, evaluating how good the graph G is to explain the occurrence of the observed data D. A kind of a standard example of such a criterion is Schwarz's[4] *Bayesian information criterion* (BIC), which is obtained by modifying (= subtracting a penalty term from) the *maximized log-likelihood* score. However, there are also many other "marginal likelihood" criteria, also named *Bayesian scores*, that are motivated by a Bayesian viewpoint; see Refs. 3 and 5. The learning task then consists in maximizing $G \mapsto \mathcal{Q}(G, D)$, given the data D.

The basic idea of an algebraic and geometric approach to this topic, proposed in Studený[6] and later developed by Studený, Vomlel and Hemmecke[7] , is to represent the BN structure given by an acyclic directed graph G by a certain vector u_G having integers as components, called the *standard imset* (for G). Note that the number of components of that vector is exponential in the number of (random) variables in consideration.

The point is that then every usual criterion $\mathcal{Q}$ for learning BN structure (namely, a score equivalent and additively decomposable one; see Refs. 8 and 9 for these concepts) becomes an affine function of the standard imset (= differs from a linear function by a constant). The main result in Ref. 7 says that the set of standard imsets is the set of vertices (= extreme points) of a certain polytope. This opens the way to apply efficient methods of polyhedral geometry (linear and integer programming) in this area. However, to do so one has to solve several open mathematical problems related to the above mentioned polytope, some of which were partially answered in Studený and Vomlel[10] .

In a conference contribution by Studený, Hemmecke and Lindner[11] an idea of an affine transformation was presented, which leads to an alternative vector representative of the BN structure, called the *characteristic imset*. Its main advantage is that it is always a zero-one vector. Moreover, we have found recently that Jaakkola, Sontag, Globerson and Meila[12] also came with the idea of application of methods of linear programming (combined with machine learning approaches) to learning BN structure.

The goal of this paper is to make an overview of our recent results concerning this topic (from recently published or submitted papers) and relate them also to the approach by Jaakkola *et al.*[12] . Besides that we present some further minor results, which we have not published so far.

In Sec. 2 we recall basic concepts. Section 3 gives the summary of recent results. Section 4 is particularly devoted to the concept of characteristic imset; besides recalling basic results on it from Hemmecke, Lindner and Studený[13] we prove here a few additional useful facts. In Sec. 5 we show that the concept of a characteristic imset allows one to give simple and elegant proofs of some (formerly known) complexity results on learning special classes of BN structures. Section 6 brings a few new results on the concept of *geometric neighborhood* for BN structures, which was introduced in Ref. 7. Section 7 contains notes about our recent computational experiments. In Conclusions we outline our future research plans.

2. Basic concepts

We tacitly assume that the reader is familiar with basic concepts from polyhedral geometry. We only recall briefly the definitions of concepts mentioned above, but skip their statistical motivation.

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 in this paper have a finite non-empty set of nodes N and two types of edges: directed edges, called *arrows* or *arcs*, denoted by $i \to j$ respectively $j \leftarrow i$, and *undirected edges*. No loops or multiple edges between two nodes are allowed.

A graph is *undirected* if all its edges are undirected. Given a graph G, its *underlying graph* $\bar{G}$ is an undirected graph obtained from G by the removal of the directions of arrows. A graph is *directed* if all its edges are arrows. A directed graph is *acyclic* if it has no directed cycle.

The set of *parents* of a node i, denoted by $\mathrm{pa}_G(i)$, is the set of nodes $j \in N$ such that $j \to i$ in G. 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. A set of nodes $C \subseteq N$ is a *clique* (or a *complete set*) in G if every pair of distinct nodes in C is connected by an undirected edge. The *degree* $\deg_G(i)$ of a node $i \in N$ in an undirected graph G is the number of nodes adjacent to i in G.

An undirected graph is called *chordal*, if every (undirected) cycle of length at least four has a chord, that is, an edge connecting two non-

consecutive nodes in the cycle. Note that an undirected graph is chordal if and only if all its edges can be directed in such a way that the result is an acyclic directed graph without immoralities; see Sec. 2.1 in Lauritzen[1] . A *forest* is an undirected graph without undirected cycles. A connected forest over N is called a *spanning tree*.

2.2. *Learning Bayesian network structure*

In statistical context, to 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 *Bayesian network* (BN) *structure* ascribed to 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. Here P is *Markovian* with respect to G if it satisfies conditional independence restrictions determined by the respective separation criterion; see Lauritzen[1] or Pearl[2] .

Different acyclic directed graphs over N could be *Markov equivalent*, which means they define the same BN structure. The classic graphical characterization of (Markov) equivalent acyclic directed graphs, provided independently by Frydenberg[14] and Verma and Pearl[15] , says that they are equivalent if and only if they have the same underlying graph and the same immoralities; for the proof see Andersson, Madigan and Perlman[16] .

The classic unique graphical representative of a BN structure is the *essential graph* G^* of the respective (Markov) equivalence class $\mathcal{G}$ of acyclic directed graphs: one has $a \to b$ in G^* if this arrow occurs in every graph from $\mathcal{G}$ and it has an undirected edge between a and b in G^* if one has $a \to b$ in one graph and $b \to a$ in another graph (from $\mathcal{G}$).

Another (unique) representative is the *pattern* $\mathrm{pat}(G)$ of arbitrary G in $\mathcal{G}$, which is obtained from the underlying graph of G by directing (only) those edges that belong to immoralities (in G). It is a less informative representative than the essential graph because there could be arrows in the essential graph which do not belong to any immorality (= are not arrows in the pattern).

Learning a BN structure means (the task) to determine it on the basis of an observed (complete) database D (of length $\ell \geq 1$), which is a sequence $x_1, \ldots, x_\ell$ of elements of the joint sample space; the database D is called *complete* if all components of the elements $x_1, \ldots, x_\ell$ are known. A *quality criterion* is a real function $\mathcal{Q}$ of two variables: of an acyclic directed graph G and of a database D. The learning procedure consists in

maximizing the function $G \mapsto \mathcal{Q}(G, D)$ for given fixed D. Since the aim is to learn a BN structure, the criterion should be *score equivalent*, which means, $\mathcal{Q}(G, D) = \mathcal{Q}(H, D)$ for any pair of Markov equivalent graphs G, H and for any database D; see Bouckaert[8] . A standard technical requirement, see Chickering[9] , is that the criterion has to be (additively) *decomposable*, which means, it can be written as follows:

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

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

Finally, let us remark that the essential graph G^* of an acyclic directed graph G is an undirected graph if and only if G has no immorality. This allows one to show that an undirected graph is the essential graph (for a class of Markov equivalent acyclic directed graphs) if and only if it is chordal. Hence, *learning decomposable models*, which are undirected graphical models ascribed to chordal graphs, see Lauritzen[1] , can be viewed as learning BN structure with the restriction to (chordal) undirected essential graphs.

2.3. *Algebraic approach to learning*

An *imset* over N is a vector in $\mathbb{Z}^{2^{|N|}}$, whose components are indexed by subsets of N. Traditionally, all subsets of N were considered, although in Sec. 4 we also consider imsets with a restricted domain.

To emphasize that components of considered vectors are indexed by subsets of N we will use notation like $\mathbb{R}^{\mathcal{P}(N)}$ or $\mathbb{Z}^{\mathcal{P}(N)}$, where $\mathcal{P}(N) \equiv \{A;\ A \subseteq N\}$ is the power set of N. Every vector in $\mathbb{R}^{\mathcal{P}(N)}$ can be written as a (real) combination of basic vectors $\delta_A \in \{0, 1\}^{\mathcal{P}(N)}$:

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

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

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

where the basic vectors can cancel each other. The standard imset is a unique algebraic representative of the corresponding BN structure because $\mathsf{u}_G = \mathsf{u}_H$ if and only if G and H are Markov equivalent; see Corollary 7.1 in Studený[6] . The convex hull of the set of all standard imsets over N is the

standard imset polytope, denoted below by P. The main result in Studený, Vomlel and Hemmecke[7] is that none of the standard imsets is a convex combination of others; thus, they are vertices of P.

A special case of the standard imset is the *elementary imset*

$$\mathsf{u}_{\langle a,b|C\rangle} := \delta_{\{a,b\}\cup C}+\delta_C-\delta_{\{a\}\cup C}-\delta_{\{b\}\cup C}, \quad a,b\in N,\ a\neq b,\ C\subseteq N\setminus\{a,b\},$$

encoding an elementary conditional independence statement $a \perp\!\!\!\perp b\,|\,C$, meaning that the variables a and b are independent conditionally the set of variables C. Indeed, one has $\mathsf{u}_{\langle a,b|C\rangle} = \mathsf{u}_G$ for an acyclic directed graph G which has only one missing edge between a and b and satisfies $\mathrm{pa}_G(a) = \mathrm{pa}_G(b) = C$. The cone E in $\mathbb{R}^{\mathcal{P}(N)}$ spanned by elementary imsets plays an important role in the algebraic approach to the description of conditional independence structures; see Studený[6] . The imsets within E describe the conditional independence structures and any standard imset belongs to the cone E, too.

An important result from the point of view of an algebraic approach to learning BN structure is that any score equivalent and decomposable quality criterion (= score function) $\mathcal{Q}$ has the form

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

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 chosen quality criterion; see Lemmas 8.3 and 8.7 in Studený[6] . The vector $t_D^{\mathcal{Q}}$ is named the *data vector* (relative to $\mathcal{Q}$). Note that the formulas for the data vector relative to the BIC and the "marginal likelihood" criteria are available; see Refs. 6 and 5.

In particular, the task to maximize $\mathcal{Q}$ is equivalent to finding the optimum of a linear function over the standard imset polytope.

3. Summary of recent results

The above optimization problem has been treated thoroughly within the *linear programming* (LP) community. The intention to apply LP methods in the area of learning BN structure motivated several open mathematical questions concerning the standard imset polytope P.

3.1. *Towards the outer description of the polytope*

A standard tool to solve LP problems is the *simplex method*; see Schrijver[17] . In order to apply the (classic) simplex method, one needs an explicit *outer*

description of the polytope via finitely many linear inequalities, that is, the characterization in the form of a polyhedron.

As concerns the standard imset polytope P, for $|N| = 3$ and $|N| = 4$ a minimal such system has 13 and 154 inequalities, respectively. However, it is already a challenge to existing software packages to find such a minimal inequality description of P for $|N| = 5$ (given by 8782 vertices). Thus, for general $|N|$, the only hope is a good guess (= conjecture about) what is the outer description of P in general.

One of our research directions was to (try to) classify necessary linear inequality constraints on P. In Studený and Vomlel[10] the case $|N| = 4$ was analyzed and the constraints characterizing $\mathsf{u} \in \mathsf{P}$ in this case were classified into three classes, namely:

- $|N| + 1$ trivial *equality constraints* of the form

$$\sum_{T \subseteq N} \mathsf{u}(T) = 0, \qquad \forall\, j \in N \quad \sum_{T \subseteq N:\, j \in T} \mathsf{u}(T) = 0\,. \tag{4}$$

- Inequality constraints that correspond to (standardized) extreme supermodular functions. Because these inequalities are valid for any vector in the cone E spanned by elementary imsets, not just for standard imsets, they were named *non-specific inequality constraints*. They have the form

$$\langle m, \mathsf{u} \rangle \equiv \sum_{T \subseteq N} m(T) \cdot \mathsf{u}(T) \geq 0\,, \tag{5}$$

  where m is a (representative on an extreme standardized) supermodular function. Here, by a *supermodular function* is meant a real function m on $\mathcal{P}(N)$ ($\equiv$ a vector in $\mathbb{R}^{\mathcal{P}(N)}$) such that

$$m(E \cup F) + m(E \cap F) \geq m(E) + m(F) \quad \text{for every } E, F \subseteq N.$$

  It is *standardized* if $m(T) = 0$ whenever $|T| \leq 1$.
- Inequality constraints corresponding to classes $\emptyset \neq \mathcal{A} \subseteq \mathcal{P}(N)$ of sets that are closed under supersets, which have the form

$$\sum_{T \in \mathcal{A}} \mathsf{u}(T) \leq 1\,. \tag{6}$$

  Because they are valid specifically only for standard imsets, they were named *specific inequality constraints*.

Note that the set of standardized supermodular functions is a pointed rational polyhedral cone in $\mathbb{R}^{\mathcal{P}(N)}$, and, therefore, has finitely many extreme

rays. Thus, (5) gives in fact only finitely many linear inequality constraints on $\mathsf{u} \in \mathsf{P}$. The problem is that one has to compute those extreme rays, which is a difficult computational task; their representatives in case $|N| \leq 5$ were computed in Ref. 18.

We conjecture that the above constraints already characterize the polytope P and the current task is either to confirm or disprove this conjecture for $|N| = 5$. Nevertheless, even if the conjecture is confirmed for general $|N|$ it only gives an *implicit* polyhedral description of P because one has to compute/characterize the extreme supermodular functions and specify all classes of subsets of N closed under supersets.

A weaker version of the conjecture was that the only lattice points in the polyhedron specified by those inequalities are the standard imsets. Actually, this weaker version of the conjecture has recently been confirmed; see Sec. 3.4.

3.2. *Geometric neighborhood*

One of possible interpretations of the simplex method is that it is a kind of search method, in which one moves between vertices of the polytope along its edges (in the geometric sense) until an optimal vertex is reached. This motivated in Studený, Vomlel and Hemmecke[7] the concept of the *geometric neighborhood* for standard imsets, and, consequently, for BN structures, or even, for particular acyclic directed graphs.

Specifically, two standard imsets are called *geometric neighbors* if the line segment connecting them is a 2-dimensional face (= a geometric edge) of the polytope P. Another research direction was to compute the geometric neighborhood for a small number of variables and (try to) interpret it.

We succeeded to compute the geometric neighborhood for $|N| = 3, 4, 5$. As a by-product we compared it for $|N| = 3$ with the well-known *inclusion neighborhood*, see Neapolitan[3] , which is at the core of current machine learning search techniques (for maximization of a quality criterion), like the so-called *GES algorithm*; see Chickering[9] . It was shown in Ref. 7 that the inclusion neighborhood is always contained in the geometric one.

Our computations suggest that, for most standard imsets, there are many more geometric neighbors than the inclusion neighbors. This observation has a simple but notable consequence from the statistical point of view: the GES algorithm may fail to find the global maximum of the quality criterion. Actually, we think that this is an inevitable defect of the inclusion

neighborhood, which may occur whenever a special strong statistical *data faithfulness assumption* is not guaranteed; for details see Ref. 7.

The result of our analysis of the geometric neighborhood in the case $|N| = 4$ was an electronic catalogue of types of geometric neighbors in Ref. 19, meant as a step towards a deeper analysis of the geometric neighborhood. We would like to find out whether one can describe geometric neighborhood in graphical terms. For further recent findings see Sec. 6.

3.3. *Lattice points in the polytope and affine transformation*

We were interested in the question of how "thick" the standard imset polytope P is. Therefore, R. Hemmecke made some computations to find out whether there exists a lattice point in its interior for $|N| \leq 5$ and the answer to this question was negative.

This led to a conjecture that every lattice point in the standard imset polytope is already a standard imset. In Ref. 10 this conjecture was confirmed. The original proof of this result in the manuscript of that paper was quite long and complicated. Later discussions among the authors of the present paper led to a much simpler proof, which was then also used in the final version of Ref. 10.

The key idea is to apply certain one-to-one linear transformation which ascribes lattice points to lattice points. The point is that the images of standard imsets are vectors, whose components are zeros and ones. As there is no lattice point in the interior of the zero-one hypercube, the above statement is immediate; we repeat the proof in Sec. 4, see Corollary 4.1.

In Studený, Hemmecke and Lindner[11] we observed that further modification of that linear transformation is useful. Specifically, we may subtract the result of the linear transformation from the constant 1-vector and get an affine transformation. The image of the standard imset u_G by that affine transformation, called the *characteristic imset* (for G) and denoted by c_G, then appears in some aspects to be even better algebraic BN structure representative than the standard imset.

In Hemmecke, Lindner and Studený[13] we show that the characteristic imsets have many elegant properties, suitable for intended application of integer programming methods to learning BN structure. They are also much closer to the graphical description than standard imsets. Section 4 recalls these arguments and adds some additional results.

3.4. *Integer programming approach*

The idea of the application of methods of *integer programming* (IP), see Schrijver[17] , is to re-formulate the task as an IP problem, that is, the task to optimize a linear function over the lattice points within a polyhedron. To this end one only needs to find a good polyhedral *LP relaxation* of the polytope of our interest, which is a polyhedron containing the polytope such that the lattice points within the polyhedron and the polytope coincide. This is the way to avoid full polyhedral description of the polytope.

Actually, in Lindner[20] an LP relaxation of the *characteristic imset polytope*, defined as the convex hull of the set of all characteristic imsets, has been suggested. Unlike the conjectured implicit polyhedral approximation of the standard imset polytope P mentioned in Sec. 3.1, this LP relaxation is *explicit*, which means all inequalities are completely specified and ready to be applied (for arbitrary $|N|$).

Jaakkola, Sontag, Globerson and Meila[12] have recently came with a slightly different idea of how to apply the methods of linear and integer programming to learning BN structures. They have used a straightforward zero-one encoding of acyclic directed graphs and transformed the task of maximizing the quality criterion to an IP problem, too. Nevertheless, they combined the LP approach with various heuristic simplifications and other machine learning methods.

The components of their vector-codes 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:

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

Moreover, they have provided an explicit polyhedral LP relaxation of their polytope, defined as the convex hull of the set of vectors $\boldsymbol{\eta}_G$ for acyclic directed graphs.

Specifically, besides elementary *non-negativity constraints* $\eta(i|B) \geq 0$ for $i \in N$, $B \subseteq N \setminus \{i\}$ and *equality constraints* $\sum_{B \subseteq N \setminus \{j\}} \eta(j|B) = 1$ for $j \in N$ they introduced so-called *cluster inequalities*, which correspond to sets $C \subseteq N$, $|C| \geq 2$:

$$1 \leq \sum_{i \in C} \sum_{D \subseteq N \setminus C} \eta(i|D)\,. \tag{7}$$

These inequalities somehow encode acyclicity requirements on the graph G.

In Studený and Haws[21] we have compared both approaches. We showed that there exists a many-to-one affine map transforming $\boldsymbol{\eta}_G$ to the standard imset u_G. The characteristic imset c_G is even a linear function of $\boldsymbol{\eta}_G$:

$$\mathsf{c}_G(S) = \sum_{i \in S} \sum_{B:\, S\setminus\{i\} \subseteq B \subseteq N\setminus\{i\}} \eta_G(i|B) \qquad \text{for } \emptyset \neq S \subseteq N. \tag{8}$$

We also succeeded to transform the inequalities provided by Jaakkola *et al.* to the framework of standard/characteristic imsets and compared the transformed inequalities with the inequalities from Sec. 3.1.

The elementary non-negativity and equality constrains for $\boldsymbol{\eta}_G$ are transformed (exactly) to the specific inequality constrains for u_G. The reader may be surprised by the fact that the transformation raises the number of inequalities. This is because of the many-to-one transformation; the fact we have to wrestle with more inequalities describing basically the same restriction is the price for having unique BN structure representatives.

The remaining cluster inequalities correspond to some of the non-specific inequality constraints. Consequently, the implicit polyhedral approximation of the standard imset polytope given by (4)-(6) is a tighter approximation of P than the transformed explicit polyhedral approximation. An interesting fact is that the some of the basic constraints $\mathsf{c}_G(S) \leq 1$ for the characteristic imsets are not implied by the transformed inequalities for $\boldsymbol{\eta}_G$.

Another non-trivial important observation from Ref. 21, which will be a basis of a future journal paper submission, is that the transformed linear inequalities define an LP relaxation of the standard/characteristic imset polytope. In particular, because the polyhedron specified by (4)-(6) is contained in this (transformed) LP relaxation of the standard imset polytope, it is also an LP relaxation of P. Thus, the weaker version of the conjecture from Sec. 3.1 has been confirmed.

Because in Ref. 20 another LP relaxation of the characteristic imset polytope has been suggested, this opens the way to the application of advanced IP methods in this area. We hope the use of characteristic imsets in learning Bayesian networks can bring saving memory demands in comparison with the use of zero-one vector codes from Ref. 12.

4. Characteristic imsets

In this section we introduce the notion of a characteristic imset and prove some useful facts about it. Throughout the section we use a special notation

for the class of subsets of N having at least two elements:

$$\mathcal{P}_2(N) \equiv \{A;\ A \subseteq N,\ |A| \geq 2\}\,,$$

and $\mathbb{Z}^{\mathcal{P}_2(N)}$ is then the set of imsets with the domain restricted to $\mathcal{P}_2(N)$.

Definition 4.1. Given an acyclic directed graph G over N, let u_G be the standard imset for G. We introduce a vector $\mathsf{p}_G \in \mathbb{Z}^{\mathcal{P}_2(N)}$ by

$$\mathsf{p}_G(S) := \sum_{X \subseteq N:\ S \subseteq X} \mathsf{u}_G(X) \qquad \text{for}\ \ S \subseteq N,\ |S| \geq 2,$$

and call it the (upper) *portrait* of u_G or, simply, of G. Moreover, the vector

$$\mathsf{c}_G := \mathbf{1} - \mathsf{p}_G \in \mathbb{Z}^{\mathcal{P}_2(N)}, \quad \text{given by } \mathsf{c}_G(S) = 1 - \mathsf{p}_G(S) \text{ for } S \in \mathcal{P}_2(N),$$

will be called the *characteristic imset* of G.

Characteristic imsets are unique representatives of Markov equivalence classes. This is because the standard imsets are unique representatives and the portrait map is a linear map that is invertible. The inverse map is given by the well-known Möbius inversion formula; see Bender and Goldman[22]. In fact, both maps assign lattice points to lattice points!

Characteristic imsets have remarkable properties and, as we will show below, their entries directly encode the underlying undirected graph and the immoralities of the given acyclic directed graph.

Theorem 4.1. *Let G be an acyclic directed graph over N. For any $S \subseteq N$, $|S| \geq 2$ we have $\mathsf{c}_G(S) \in \{0,1\}$ and $\mathsf{c}_G(S) = 1$ iff there exists some $i \in S$ with $S \setminus \{i\} \subseteq \mathrm{pa}_G(i)$. In particular, $\mathsf{c}_G \in \{0,1\}^{\mathcal{P}_2(N)}$.*

Proof. Consider the defining formula (2) for the standard imset. For any $S \subseteq N$, $|S| \geq 2$, the value $\mathsf{p}_G(S)$ can be computed as

$$\mathsf{p}_G(S) = \sum_{X \subseteq N:\ S \subseteq X} \mathsf{u}_G(X) = 1 + \sum_{i \in N:\ S \subseteq \mathrm{pa}_G(i)} 1 - \sum_{i \in N:\ S \subseteq \mathrm{pa}_G(i) \cup \{i\}} 1\,.$$

Hence, we get

$$\begin{aligned} \mathsf{c}_G(S) = 1 - \mathsf{p}_G(S) &= \sum_{i \in N:\ S \subseteq \mathrm{pa}_G(i) \cup \{i\}} 1 - \sum_{i \in N:\ S \subseteq \mathrm{pa}_G(i)} 1 \\ &= \sum_{i \in N:\ S \subseteq \mathrm{pa}_G(i) \cup \{i\}, i \in S} 1 = \sum_{i \in S:\ S \setminus \{i\} \subseteq \mathrm{pa}_G(i)} 1\,. \end{aligned}$$

For fixed S, assume that there are two different elements $i, j \in S$ with $S \setminus \{i\} \subseteq \mathrm{pa}_G(i)$ and $S \setminus \{j\} \subseteq \mathrm{pa}_G(j)$. This implies both $i \in \mathrm{pa}_G(j)$

and $j \in \mathrm{pa}_G(i)$. The simultaneous existence of the arcs $i \to j$ and $j \to i$, however, contradicts the assumption that G is acyclic. Therefore, for each $S \subseteq N$, there is at most one $i \in S$ with $S \setminus \{i\} \subseteq \mathrm{pa}_G(i)$. Consequently,

$$\mathsf{c}_G(S) = \sum_{i \in S:\ S \setminus \{i\} \subseteq \mathrm{pa}_G(i)} 1 \quad \in \{0,1\},$$

and, thus, $\mathsf{c}_G \in \{0,1\}^{\mathcal{P}_2(N)}$. □

Corollary 4.1. *For any N, the only lattice points in the standard imset polytope and in the characteristic imset polytope are their vertices.*

Proof. The statement holds for any zero-one polytope and thus, in particular, also for the characteristic imset polytope. Moreover, the portrait map and its inverse, the Möbius map, are linear mappings between u_G and c_G that map lattice points to lattice points. Thus, the result holds also for the standard imset polytope. □

Given a chordal undirected graph H, the corresponding characteristic imset c_H can be introduced as the characteristic imset of any acyclic directed graph G, whose essential graph is H. The observation that characteristic imsets are unique representatives of Markov equivalence classes makes the definition correct.

Corollary 4.2. *Let H be an undirected chordal graph over N. Then, for $S \subseteq N$, $|S| \geq 2$, we have $\mathsf{c}_H(S) = 1$ if and only if S is a clique in H.*

Proof. As H is the essential graph of an acyclic directed graph K which has no immorality, we can direct the edges of H in such a way that we obtain an equivalent acyclic directed graph G without an immorality. To show the forward implication, let $S \subseteq N$, $|S| \geq 2$ be given with $\mathsf{c}_H(S) = 1$. As $\mathsf{c}_H(S) = \mathsf{c}_G(S) = 1$, there exists some $i \in S$ such that $S \setminus \{i\} \subseteq \mathrm{pa}_G(i)$. Assume now, for a contradiction, that there are two nodes $j, k \in S \setminus \{i\}$ that are not adjacent by an edge in G (and hence j and k are not adjacent in H). Then, however, $j \to i \leftarrow k$ is an immorality in G, a contradiction. Hence, all nodes in $S \setminus \{i\}$ must be pairwise connected by an edge in H. As they are all connected in H by an edge to i, S is a clique in H.

To show the converse, let $S \subseteq N$ be a clique in H. Note that in G, being an acyclic directed graph, the set S must contain a node i such that, for all $j \in S \setminus \{i\}$, the edge between i and j in H is directed towards i in G. But then $S \setminus \{i\} \subseteq \mathrm{pa}_G(i)$ and, therefore, $\mathsf{c}_H(S) = 1$ by Theorem 4.1. □

Applying this observation to special undirected chordal graphs, namely to undirected forests, we obtain the following characterization.

Corollary 4.3. *Let H be an undirected forest having N as the set of nodes. Then, for $S \subseteq N$, $|S| \geq 2$, we have $\mathsf{c}_H(S) = 1$ if and only if S is an edge in H, or, in other words,*

$$\mathsf{c}_H = \begin{pmatrix} \chi(H) \\ \mathbf{0} \end{pmatrix},$$

where $\chi(H)$ denotes the characteristic vector of the edge-set of H.

Indeed, the only cliques of cardinality at least two in a forest are its edges. A similar result, in fact, holds for any acyclic directed graph G.

Corollary 4.4. *Let G be an acyclic directed graph over N and $\bar{G}$ its underlying undirected graph. Then for any two-element subset $\{a, b\} \subseteq N$, we have $\mathsf{c}_G(\{a, b\}) = 1$ if and only if $a \to b$ or $b \to a$ is an edge in G, or, in other words,*

$$\mathsf{c}_G = \begin{pmatrix} \chi(\bar{G}) \\ \star \end{pmatrix},$$

where $\star$ denotes the remaining components of c_G.

Proof. This is an easy consequence of Theorem 4.1. If $\mathsf{c}_G(S) = 1$ for $S = \{a, b\}$ then the only $i \in S$ with $S \setminus \{i\} \subseteq \mathrm{pa}_G(i)$ are either a or b. □

Thus, c_G is an extension of the characteristic vector $\chi(\bar{G})$ of the edge-set of $\bar{G}$, which motivated our terminology. Let us now show how to convert c_G back to the pattern graph $\mathrm{pat}(G)$ of G.

Theorem 4.2. *Let G be an acyclic directed graph over N and $a, b \in N$ are distinct nodes. Then the following holds:*

(i) *$a, b \in N$ are adjacent (= connected by an edge) in G if and only if $\mathsf{c}_G(\{a, b\}) = 1$, otherwise $\mathsf{c}_G(\{a, b\}) = 0$.*

(ii) *$a \to b$ belongs to an immorality in G if and only if there exists some $i \in N \setminus \{a, b\}$ with $\mathsf{c}_G(\{a, b, i\}) = 1$ and $\mathsf{c}_G(\{a, i\}) = 0$. The latter condition implies $\mathsf{c}_G(\{a, b\}) = 1$ and $\mathsf{c}_G(\{b, i\}) = 1$.*

Proof. The condition *(i)* follows from Corollary 4.4. For *(ii)* assume that $a \to b \leftarrow i$ is an immorality in G. Then $\mathsf{c}_G(\{a, b, i\}) = 1$ by Theorem 4.1 and the necessity of the other conditions follows from *(i)*. Conversely, provided

that $\mathsf{c}_G(\{a,b,i\}) = 1$, one of the three options $a \to i \leftarrow b$, $i \to a \leftarrow b$ and $a \to b \leftarrow i$ (with possible additional edges) occurs. Now, $\mathsf{c}_G(\{a,i\}) = 0$ implies that a and i are not adjacent in G, which excludes the first two options and implies $a \to b \leftarrow i$ must be an immorality. □

Corollary 4.5. *Given an acyclic directed graph G, the characteristic imset c_G is determined uniquely by its values for sets of cardinality 2 and 3.*

Proof. By Theorem 4.2 these values determine both the underlying graph and immoralities in G. In particular, they determine the pattern $\mathrm{pat}(G)$. As explained in Sec. 2.2, this uniquely determines the BN structure and, therefore, the respective standard and characteristic imsets. □

More specifically, the components of c_G for $|S| \geq 4$ can be derived iteratively from the components for $|S| \leq 3$ on the basis of the following lemma. A further simple consequence of the lemma below is that the entries for $|S| \geq 4$ are not linear functions of the entries for $|S| \leq 3$.

Lemma 4.1. *Let G be an acyclic directed graph over N, and $S \subseteq N$, $|S| \geq 4$. Then the following conditions are equivalent.*

(a) $\mathsf{c}_G(S) = 1$,
(b) there exist $|S| - 1$ *subsets* T *of* S *with* $|T| = |S| - 1$ *and* $\mathsf{c}_G(T) = 1$,
(c) there exist three subsets T *of* S *with* $|T| = |S| - 1$ *and* $\mathsf{c}_G(T) = 1$.

In the proof, by a *terminal node* within a set $T \subseteq N$ we mean $i \in T$ such that there is no $j \in T \setminus \{i\}$ with $i \to j$ in G.

Proof. The implication $(a) \Rightarrow (b)$ follows from Theorem 4.1; $(b) \Rightarrow (c)$ is trivial. To show $(c) \Rightarrow (a)$ we first fix a terminal node i within S. Now, (c) implies there exist at least two sets $T \subseteq S$, $|T| = |S| - 1$ which contain i. Let $\tilde{T}$ be one of them. Since $\mathsf{c}_G(\tilde{T}) = 1$ by Theorem 4.1, there exists $k \in \tilde{T}$ with $j \to k$ for every $j \in \tilde{T} \setminus \{k\}$. If $i \neq k$, then $i \to k$, which contradicts i to be terminal in S. Thus, $i = k$. Since, those two sets T cover S one has $j \to i$ for every $j \in S \setminus \{i\}$ and Theorem 4.1 implies $\mathsf{c}_G(S) = 1$. □

Theorem 4.2 allows us to reconstruct the essential graph for G. Indeed, the conditions *(i)* and *(ii)* directly characterize the pattern graph $\mathrm{pat}(G)$. However, in general, there could be other arrows in the essential graph. Fortunately, there is a polynomial graphical algorithm transforming $\mathrm{pat}(G)$ into the corresponding essential graph G^*. More specifically, Theorem 3 in

Meek[23] says that provided pat(G) is the pattern of an acyclic directed graph G the repeated (exhaustive) application of the orientation rules from Figure 1 gives the essential graph G^*.

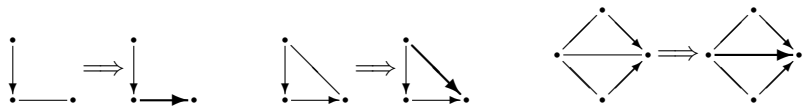

Fig. 1. Orientation rules for getting the essential graph.

Finally, we wish to point out that Theorems 4.1 and 4.2 directly lead to a procedure for testing whether a given vector $\mathsf{c} \in \mathbb{Z}^{\mathcal{P}_2(N)}$ is a characteristic imset for some (acyclic directed graph) G over N. Using both theorems, one first constructs a candidate pattern graph, then a candidate essential graph, and then from it a candidate acyclic directed graph G. It remains to check whether the characteristic imset of G coincides with the given vector c.

5. Complexity of learning (special) chordal graphs

A lot of research effort in machine learning community has been devoted to deriving complexity results on learning BN structure, analyzing different optimization strategies, scoring functions and representations of data. For example, Chickering, Heckerman and Meek[24] showed the large-sample learning problem to be NP-hard even when the distribution is perfectly Markovian. Similarly, Chickering[25] showed learning BN structure to be NP-complete when using a certain Bayesian score. This remains to be valid even if the number of parents is limited to a constant.

In this section, we deal with learning decomposable models interpreted as BN structures, see Sec. 2.2, and also derive some complexity results. What is special here is that we use as a tool for derivation of our complexity results *characteristic imsets*, introduced in Sec. 4. The point is that the characteristic imsets bring a clear insight into the learning task, and, thus, simplify the complexity considerations. We do not claim that the complexity observations themselves are strikingly new; we more likely offer elegant and simple proofs in comparison with the original machine learning treatments.

Our assumptions

The input to the learning problems we consider is a prescribed undirected graph K over N and an evaluation oracle for the score function $\mathcal{Q}$. The goal of the learning problem is then to learn an acyclic directed graph G over N that maximizes the quality criterion and whose essential graph is an (undirected) subgraph of K of a certain type. In particular, we are interested in learning undirected forests and spanning trees with and without degree bounds and learning undirected chordal graphs.

We wish to point out here that we make minimal assumptions on the database D and on the quality criterion to be optimized. We only assume that the database D over N is *complete* and the quality criterion (= score function) we require to be *score equivalent* and *decomposable*.

In fact, instead of having D and an explicit score function available, we only assume that we are given an evaluation oracle (depending on D) that, when queried on G, returns the value $\mathcal{Q}(G, D)$ of the quality criterion. Clearly, especially for larger databases D, computing a single score function value $\mathcal{Q}(G, D)$ may be expensive. By assuming we are given an evaluation oracle, we give a constant cost to score function evaluations in our complexity results below. It also implicitly means that the (large or small) number of data items in D will be irrelevant for our complexity considerations.

Finally, we remind the reader that under our assumptions the learning of the best acyclic directed graph G for D becomes the problem of maximizing a certain *linear functional*, depending on D only, over the set of characteristic imsets; see Sec. 2.3 and Sec. 4.

5.1. *Learning undirected forests and spanning trees*

By Corollary 4.3, we know that every acyclic directed graph whose essential graph is an undirected forest H has $\binom{\chi(H)}{\mathbf{0}}$ as its characteristic imset. Thus, the problem of learning the best undirected forest is equivalent to maximizing a linear functional over such vectors $\binom{\chi(H)}{\mathbf{0}}$ which in turn is equivalent to finding a maximum weight forest H as a subgraph of K. The same arguments hold for learning undirected spanning subtrees of K. These are two well-known combinatorial problems that can be solved in polynomial time via greedy-type algorithms; see, for example, Sec. 40 in Schrijver[26]. We conclude the following statement.

Lemma 5.1. *Given an undirected graph $K = (N, \mathcal{E}(K))$ and an evaluation oracle for computing $\mathcal{Q}$, the problems of finding a maximum score subgraph*

of K that is

(a) a forest,
(b) a spanning tree,

can be solved in time polynomial in $|N|$.

Although K is a part of the input, we need not state the complexity dependence with respect to the encoding length of K explicitly here, since the encoding length $\langle K \rangle$ is at least $|N|$. Moreover, we have $\langle K \rangle \in O(|N|^2)$.

Chow and Liu[27] provided a polynomial time procedure (in $|N|$) for maximizing the maximized log-likelihood score which finds an optimal dependence tree (= a spanning tree). The core of their algorithm is the greedy algorithm and they apply it to a non-negative objective function. For their result, the complexity of computing the probabilities from data (and hence the objective/score function) is also omitted. A similar result was obtained by Heckerman, Geiger and Chickering[28] for the Bayesian scores. Our result combines all of these previous results by only supposing a decomposable and score equivalent quality criterion.

We wish to point out here that the well-known GES algorithm, see Refs. 29 and 9, designed to learn general BN structure, could be modified in a straightforward way to learn undirected forests (among the subgraphs of K). Then the first phase of this new GES-type algorithm coincides with the greedy algorithm to find a maximum weight forest, the second phase of the algorithm cannot remove any edge. Thus, the modified GES algorithm always finds a best undirected forest in time polynomial in $|N|$.

5.2. *Learning forests and trees with a degree bound*

Although the problems of learning undirected forests and spanning trees are solvable in polynomial time, learning an undirected forest/spanning tree with a given degree bound $\deg_G(i) \leq k$ for any $i \in N$, where $2 \leq k < |N|-1$, is NP-hard. For $k = 1$ this problem coincides with the well-known problem of finding a maximum weight matching in K, which is, in the general case, solvable in polynomial time; see Sec. 30 in Schrijver[26] . However, for $k \geq 2$, the situation is different.

Theorem 5.1. *Given an undirected graph $K = (N, \mathcal{E}(K))$, an evaluation oracle for computing $\mathcal{Q}$ and a constant $k \in \mathbb{Z}$ with $2 \leq k < |N| - 1$, the following statements hold.*

(i) The problem of finding a maximum score subgraph of K that is a forest

and fulfils the degree bounds $\deg(i) \leq k$, $\forall i \in N$, *is NP-hard in* $|N|$ *for any fixed score function* $\mathcal{Q}$ *representable by a (strictly) positive vector.*

(ii) The problem of finding a maximum score spanning tree of K *that fulfils the degree bounds* $\deg(i) \leq k$, $\forall i \in N$, *is NP-hard in* $|N|$ *for any fixed score function* $\mathcal{Q}$.

As $\langle K \rangle \in O(|N|^2)$, we have again omitted the explicit dependence on $\langle K \rangle$.

Proof. We deduce part *(ii)* from the following feasibility problem. In Sec. 3.2.1 of Garey and Johnson[30], the following task has been shown to be NP-complete by the reduction to the Hamiltonian path problem:

> Bounded degree spanning tree
> Instance: An undirected graph K and a constant $2 \leq k < |N| - 1$.
> Question: Is there a spanning tree for K in which no node has degree exceeding k?

Part *(i)* now follows by considering the subfamily of problems in which the respective linear objective $\chi(H) \mapsto \mathbf{q}^\intercal \chi(H)$ (representing $\mathcal{Q}$) is given by a vector $\mathbf{q}$ with (only strictly) positive components and, thus, every optimal forest (with the bounded degree) is a spanning tree. Hence, the problem of finding a maximum-weight forest (with a given degree bound) is equivalent to finding a maximum-weight spanning tree (with a given degree bound). As the feasibility problem for the latter is NP-complete, part *(i)* follows. □

We wish to remark that Meek[31] showed a similar hardness result for learning paths, that is, spanning trees with upper degree bound $k = 2$ for the maximized log-likelihood score, BIC and Bayesian scores.

5.3. *Learning chordal graphs*

Undirected chordal graph models are the intersection of Bayesian network models and undirected graph models, known as Markov networks; see Sec. 3.4.1 in Ref. 6. Here, we show that learning these models is NP-hard.

Theorem 5.2. *Given an undirected graph* $K = (N, \mathcal{E}(K))$ *and an evaluation oracle for computing* $\mathcal{Q}$, *the problem of finding a maximum score chordal subgraph of* K *is NP-hard in* $|N|$.

Proof. We show that one can polynomially transform the following NP-hard problem to learning undirected chordal graphs:

Clique of given size
Instance: An undirected graph K and a constant $2 \leq k \leq |N| - 1$.
Question: Is there a clique set in K of size at least k?

To this end we define a learning problem that would solve this problem. By Corollary 4.2 we know that, for any chordal graph G, the entry $\mathsf{c}_G(S)$ is 1 iff $S \subseteq N$ is a clique; otherwise this entry is 0. Thus, the score function value for G is determined by the values of the linear objective function $\mathsf{c} \mapsto \mathbf{q}^\intercal \mathsf{c}$ for the cliques S in G. In particular, we can define the values for the cliques in such a way that when transforming the learning problem to the problem of maximizing $\mathbf{q}^\intercal \mathsf{c}$ over the characteristic imset polytope, the entries $\mathbf{q}(S)$ are 0 when $|S| < k$ and positive when $|S| \geq k$. This implies that the maximum score among all chordal subgraphs of K is positive iff there exists a chordal subgraph in K containing a clique S of size $|S| \geq k$. This happens iff K has a clique of size at least k. □

5.4. *Learning chordal graphs with bounded size of cliques*

Let us consider a variation of the previous task by introducing an upper bound ℓ for the size of cliques. If $\ell \leq 2$, we get the problems of learning undirected forests/matchings, which we already know are solvable in polynomial time; see Secs. 5.1 and 5.2.

For $\ell > 2$, the corresponding problem is NP-hard already for a fixed type of score function. This has been shown by Srebro[32] for the maximized log-likelihood score, as a generalization of the work by Chow and Liu[27] .

6. Geometric neighborhood

In this section, we present a few facts concerning the geometric neighborhood, introduced in Sec. 3.2 for BN structures, respectively for acyclic directed graphs over N. Specifically, a pair of geometric neighbors corresponds to a 2-face (= a geometric edge) of the standard imset polytope. This allows us to derive the following characterization of geometric neighbors of a *full acyclic directed graph* over N. This is any acyclic directed graph over N, whose underlying undirected graph is complete (= N is a clique in the underlying graph).

Theorem 6.1. *An acyclic directed graph G over N is a geometric neighbor of a full acyclic directed graph H over N if and only if G is full with the exception of (just) one missing edge. In particular, the geometric neighbors and the inclusion neighbors of H coincide.*

Proof. The standard imset for H is the zero imset $\mathsf{u}_H \equiv 0$. First, we observe that, for any elementary imset $\mathsf{u}_{\langle a,b|C\rangle}$ (see Sec. 2.3), the line segment in $\mathbb{R}^{\mathcal{P}(N)}$ connecting $\mathsf{u}_H = 0$ and $\mathsf{u}_{\langle a,b|C\rangle}$ is a face of P; this means, they are geometric neighbors. The observation follows easily from the well-known fact that elementary imsets generate the extreme rays of the polyhedral cone E spanned by elementary imsets; see, for example, Kashimura, Sei, Takemura and Tanaka[33] . Indeed, both the zero imset and every elementary imset belong to P and P is a subset of E.

The fact that there are no other 2-faces of P containing $\mathsf{u}_H = 0$ is also a consequence of $\mathsf{P} \subseteq \mathsf{E}$. If the line segment $[\mathsf{u}_H, \mathsf{v}]$ for $\mathsf{v} \in \mathsf{P}$ is a face of P, then v is a conic combination of elementary imsets. It cannot be a combination of more than one elementary imset, for otherwise $[\mathsf{u}_H, \mathsf{v}]$ is not a face of P (as elementary imsets also belong to P).

Thus, the geometric neighbors of u_H are elementary imsets, which correspond to (acyclic directed) graphs with just one missing edge; see Sec. 2.3. These are known to coincide with the inclusion neighbors of u_H, respectively of H; see Corollary 8.4 in Ref. 6. □

Nevertheless, as explained in Sec. 4, the standard imset polytope P is affine isomorphic to the characteristic imset polytope. In particular, the geometric neighborhood can equivalently be introduced through characteristic imsets: simply, the geometric neighbors correspond to 2-faces of the characteristic imset polytope.

This leads to a method to prove that two (non-equivalent) acyclic directed graph G and H over N are geometric neighbors, which appears to be useful in some cases. Specifically, we can do so as follows: we construct a vector $\mathbf{q} \in \mathbb{R}^{\mathcal{P}_2(N)}$ in such a way that the linear objective function $\mathsf{c} \mapsto \mathbf{q}^\intercal \mathsf{c}$ achieves its maximum over characteristic imsets for acyclic directed graphs (over N) just in the graphs that are Markov equivalent either to G or H.

Using this method we can characterize geometric neighbors of the *empty* acyclic directed graph (= the graph over N which has no edge). We show that its geometric neighbors are graphs over N that have just one *non-initial node*, that is, just one node $i \in N$ with $\mathrm{pa}_G(i) \neq \emptyset$.

Lemma 6.1. *If H is the empty graph and G a graph (over N) with just one node $a \in N$ with $\mathrm{pa}_G(a) \neq \emptyset$, then G and H are geometric neighbors.*

Proof. Let's assume $\mathrm{pa}_G(a) = \{i_1, \ldots, i_m\}$, $m \geq 1$ and put $T \equiv \{i_1, \ldots, i_m\} \cup \{a\}$. Note that for all $S \subseteq T$ such that $a \in S$ and $|S| \geq 2$,

one has $\mathsf{c}_G(S) = 1$, otherwise $\mathsf{c}_G(S) = 0$. Define $\mathbf{q} \in \mathbb{R}^{\mathcal{P}_2(N)}$ as follows:

$$\mathbf{q}(S) := \begin{cases} -1 & S \subset T,\, a \in S, \\ 2^m - 2 & S = T, \\ -2 & \text{otherwise}, \end{cases}$$

where $\subset$ denotes strict inclusion. Observe that one has just $2^m - 2 \geq 0$ sets $S \subset T$ with $a \in S$ and $|S| \geq 2$. Hence, $\mathbf{q}^\intercal \mathsf{c}_G = 0$. As $\mathsf{c}_H = 0$, $\mathbf{q}^\intercal \mathsf{c}_H = 0$.

Let K be an acyclic directed graph over N which is a maximizer of $\mathsf{c}_K \mapsto \mathbf{q}^\intercal \mathsf{c}_K$ among acyclic directed graphs over N. To show it is Markov equivalent either to H or G we distinguish two cases:

- If $\mathsf{c}_K(T) = 0$, then because all other components of $\mathbf{q}$ are negative, $\mathsf{c}_K(S) = 0$ for all $S \subseteq N$, $|S| \geq 2$. Thus, $K = H$.
- If $\mathsf{c}_K(T) = 1$, then a unique $b \in T$ exists such that $T \setminus \{b\} \subseteq \mathrm{pa}_K(b)$. Furthermore, for all $2^m - 2$ subsets $S \subset T$ with $|S| \geq 2$ and $b \in S$ it follows that $\mathsf{c}_K(S) = 1$. By the definition of $\mathbf{q}$, for every such S, $\mathbf{q}(S) \in \{-1, -2\}$. Observe that K cannot have other edges except those directed from $T \setminus \{b\}$ to b. Indeed, if K has additional edges, then we compare it with the graph $\tilde{K}$ having just the arrows from $T \setminus \{b\}$ to b and observe that $\mathbf{q}^\intercal \mathsf{c}_{\tilde{K}} > \mathbf{q}^\intercal \mathsf{c}_K$, which contradicts the assumption that K is a maximizer. Therefore, if $b = a$ then $K = G$. If $b \neq a$ but $m = 1$ then K is the graph with the only arrow $a \to b$ while G is the graph with the only arrow $b \to a$. Thus, K is Markov equivalent to G. The case $b \neq a$ and $m \geq 2$ is not possible: it implies the existence of a set $S \subset T$ with $|S| \geq 2$ and $b \in S$ such that $\mathbf{q}(S) = -2$. This means $\mathbf{q}^\intercal \mathsf{c}_K < 0$ contradicting the assumption that K is a maximizer.

Therefore, the line segment connecting c_H and c_G is a face of the characteristic imset polytope. □

Theorem 6.2. *The geometric neighbors of the empty graph H are just those (acyclic directed) graphs G that have only one non-initial node.*

Proof. By Lemma 6.1 we know all such G are geometric neighbors of H. To show that these are the only neighbors of H it suffices to show that, for any acyclic directed graph K over N, c_K can be written as a non-negative linear combination of vectors c_G for such graphs G.

We prove it by induction on the number of non-initial nodes in K. If there is only one node $i \in N$ with $\mathrm{pa}_K(i) \neq \emptyset$ then K has the form of G and we are finished. If one has $k \geq 2$ such nodes, then find a terminal node $a \in N$ in the graph K. Introduce a graph G over N whose only arrows

are the arrows directed from $\mathrm{pa}_K(a)$ to a. Let us denote by L the graph obtained from K by the removal of these arrows in G. It is easy to verify that $\mathsf{c}_K = \mathsf{c}_L + \mathsf{c}_G$. Since L has smaller number of non-initial nodes than K, by the induction hypothesis, c_L is the desired combination of characteristic imsets for geometric neighbors of H. Hence, the same conclusion for c_K can be derived. □

7. Computational experiments

This section contains a few general notes on our preliminary computational experiments based on the polyhedral approach. We plan to continue in the experiments and, when finished, to prepare a paper devoted to them.

First, we comment on some common machine learning techniques used to learn BN structure. A speedy method is the hill-climbing approach used in the GES algorithm; see Meek[29] and Chickering[9] . However, as mentioned in Sec. 3.2, this approach does not guarantee to find the global maximum of the quality criterion $\mathcal{Q}$. Silander and Myllymäki[34] offered an approach based on *dynamic programming* that already allows one to find the global maximum; de Campos, Zeng and Ji[35] were also interested in finding the global maximum and came with the idea of the use of a (general) *branch and bound* approach and the idea of the reduction of the search space, based on a more detailed analysis of the particular form of local scores.

Jaakkola *et al.*[12] already used an LP approach and also utilized the idea of the search space reduction from Ref. 35. In the context of their approach, see Sec. 3.4, this idea of *pruning* of the search space can be described as follows. Consider the criterion $\mathcal{Q}$ in the form (1), where $q_{i|B}(D_{\{i\}\cup B})$ are the *local scores*. If the database D is such that, for some $i \in N$ and $C \subset B \subseteq N \setminus \{i\}$ one has $q_{i|C}(D_{\{i\}\cup C}) > q_{i|B}(D_{\{i\}\cup B})$ then no optimal acyclic graph G over N has $\mathrm{pa}_G(i) = B$ and one can, therefore, limit oneself to codes $\boldsymbol{\eta}$ with $\eta(i|B) = 0$. If $q_{i|C}(D_{\{i\}\cup C}) \geq q_{i|B}(D_{\{i\}\cup B})$ then at least one optimal graph G satisfies $\mathrm{pa}_G(i) \neq B$. Thus, provided that the aim is to find just one optimal graph, one can also, without loss of generality, put $\eta(i|B) = 0$. This allows one to reduce the number of components of $\boldsymbol{\eta}$.

In our computational experiments, we represented BN structures by characteristic imsets. However, because of the relation (8), we can also exclude the component $\mathsf{c}_G(S)$, $S \in \mathcal{P}_2(N)$, provided one is sure that, for each $i \in S$ and B with $S \setminus \{i\} \subseteq B \subseteq N \setminus \{i\}$ the component $\eta(i|B)$ can be put to zero. Thus, we have also represented $\mathcal{Q}(*, D)$ in the form of a cache of local scores $q_{i|B}(D_{\{i\}\cup B})$, $i \in N$, $B \subseteq N \setminus \{i\}$.

The general idea was to take the LP relaxation of the characteristic imset polytope from Lindner[20] , formulate the learning task as an IP problem and use the state-of-art software like CPLEX[36] . However, three major bottlenecks appear:

(1) exponentially many variables in $|N|$,
(2) exponentially many inequalities in $|N|$,
(3) computing the objective is time consuming.

To handle these problems we used the pruning method mentioned above and combined it with the row-generation techniques and the branch-and-bound method to solve the IP, see Sec. 24.1 in Schrijver[17] . Besides learning general BN structures, we made some experiments with learning chordal graphs with a prescribed upper bound on the size of cliques.

The examples we analyzed were taken from the UCI-Machine learning repository[37] , were generated "randomly" and also taken from Ref. 12. The main observation/conclusion from our experiments is that the pruning step is crucial and reduces the total computation time tremendously.

8. Conclusions

To summarize, we offer a new method for analyzing the learning procedure through an algebraic way of representing statistical models. Characteristic imsets turn out to be very natural encodings of BN structures (= Markov equivalence classes) that are much closer to the graphical description. From the characteristic imset, the associated essential graph can easily be reconstructed, since it directly encodes the pattern of this equivalence class.

Characteristic imsets allow one to reduce the combinatorial learning task to a linear (integer) optimization problem that may/will lead to future applications of efficient (integer) linear programming methods and software in this area. Moreover, they also offer elegant combinatorial proofs for known results and allow one to establish new complexity results for learning restricted BN structures such as undirected forests or spanning trees. These proofs avoid special assumptions on the form of the quality criterion besides the standard assumptions of score equivalence and decomposability. The simplicity of these constructions gives a hope that characteristic imsets will be a very useful tool to unravel more interesting theoretical and algorithmic results for the learning of BN structures that were hidden so far due to a lack of a suitable way of encoding of BN structures.

In our future work, we plan to study further the standard imset and

the characteristic imset polytopes, respectively. We will apply tools from integer linear programming to learning BN structure. Although the linear optimization problem is defined for $2^{|N|} - |N| - 1$ variables, one can employ pruning or prescribed size restrictions in practice to reduce the optimization problem down to only a few hundreds or thousands of (integer) variables even for $|N|$ between 30 and 40. Indeed, our first preliminary computations using characteristic imsets are very promising.

Acknowledgments

The research of Milan Studený has been supported by the grant GAČR n. 201/08/0539.

References

1. S. L. Lauritzen, *Graphical Models* (Clarendon Press, 1996).
2. J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, 1988).
3. R. E. Neapolitan, *Learning Bayesian Networks* (Pearson Prentice Hall, 2004).
4. G. Schwarz, *The Annals of Statistics* **6**, 461 (1978).
5. M. Studený, Mathematical aspects of learning Bayesian networks: Bayesian quality criteria, research report n. 2234, Institute of Information Theory and Automation, Prague, December 2008.
6. M. Studený, *Probabilistic Conditional Independence Structures* (Springer Verlag, 2005).
7. M. Studený, J. Vomlel and R. Hemmecke, *International Journal of Approximate Reasoning* **51**, 578 (2010).
8. R. R. Bouckaert, Bayesian belief networks: from construction to evidence, PhD thesis, University of Utrecht, 1995.
9. D. M. Chickering, *Journal of Machine Learning Research* **3**, 507 (2002).
10. M. Studený, J. Vomlel, *International Journal of Approximate Reasoning* **52**, 627 (2011).
11. M. Studený, R. Hemmecke and S. Lindner, Characteristic imset: a simple algebraic representative of a Bayesian network structure, in Proceedings of the 5th European Workshop on Probabilistic Graphical Models, HIIT Publications, 2010, 257-264.
12. T. Jaakkola, D. Sontag, A. Globerson and M. Meila, Learning Bayesian network structure using LP relaxations, in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 2010, 358-365.
13. R. Hemmecke, S. Lindner and M. Studený, Characteristic imsets for learning Bayesian network structure, submitted to *International Journal of Approximate Reasoning*.
14. M. Frydenberg, *Scandinavian Journal of Statistics* **17**, 333 (1990).
15. T. Verma and J. Pearl, Equivalence and synthesis of causal models, in Uncertainty in Artificial Intelligence 6, Elsevier, 1991, 220-227.

16. S. A. Andersson, D. Madigan and M. D. Perlman, *The Annals of Statistics* **25**, 505 (1997).
17. A. Schrijver, *Theory of Linear and Integer Programming* (John Wiley, 1986).
18. M. Studený, R. R. Bouckaert and T. Kočka, Extreme supermodular set functions over five variables, research report n. 1977, Institute of Information Theory and Automation, Prague, January 2000.
19. J. Vomlel, M. Studený, The catalogue of types of geometric neighbors over four variables, available at staff.utia.cas.cz/vomlel/imset/catalogue-diff-imsets-4v.html
20. S. Lindner, Discrete optimisation in machine learning - learning of Bayesian network structures and conditional independence implication, PhD thesis, Technische Universität Munich, 2012.
21. M. Studený, D. Haws, On polyhedral approximations of polytopes for learning Bayes nets, research report n. 2303, Institute of Information Theory and Automation of the ASCR, Prague, July 2011, also available at http://arxiv.org/abs/1107.4708
22. E. A. Bender and J. R. Goldman, *The American Mathematical Monthly* **82**, 789 (1975).
23. C. Meek, Causal inference and causal explanation with background knowledge, in Uncertainty in Artificial Intelligence 11, Morgan Kaufmann, 1995, 403-410.
24. D. M. Chickering, D. Heckerman and C. Meek, *Journal of Machine Learning Research* **5**, 1287 (2004).
25. D. M. Chickering, Learning Bayesian networks is NP-complete, in Learning from Data: Artificial Intelligence and Statistics V, Springer Verlag, 1996, 121-130.
26. A. Schrijver, *Combinatorial Optimization - Polyhedra and Efficiency*, volume B (Springer Verlag, 2003).
27. C. K. Chow and C. N. Liu, *IEEE Transactions on Information Theory* **14**, 462 (1968).
28. D. Heckerman, D. Geiger and D. M. Chickering, *Machine Learning* **20**, 197 (1995).
29. C. Meek, Graphical models: selecting causal and statistical models, PhD thesis, Carnegie Mellon University, 1997.
30. M. R. Garey and D. S. Johnson, *Computers and Intractability - A Guide to the Theory of NP-Completeness* (Bell Telephone Laboratories, 1979).
31. C. Meek, *Journal of Artificial Intelligence Research* **15**, 383 (2001).
32. N. Srebro, Maximum likelihood bounded tree-with Markov networks, in Uncertainty in Artificial Intelligence 17, Morgan Kaufmann, 2001, 504-511.
33. T. Kashimura, T. Sei, A. Takemura and K. Tanaka, Cones of elementary imsets and supermodular functions: a review and some new results, submitted to Proceedings of 2nd CREST-SBM International Conference *Harmony of Gröbner Bases and the Modern Industrial Society*, World Scientific, 2012.
34. T. Silander and P. Myllymäki, A simple approach for finding the globally optimal Bayesian network structure, in Uncertainty in Artificial Intelligence 22, AUAI Press, 2006, 445-452.

35. C. P. de Campos, Z. Zeng and Q. Ji, Structure learning of Bayesian networks using constraints, in Proceedings of the 26th Annual International Conference on Machine Learning (ICML), 2009, 113-120.
36. Ilog team, CPLEX - mathematical programming optimizer, available electronically at www-01.ibm.com/software/integration/optimization/cplex/
37. University California Irvine (UCI) Machine learning repository, available at http://archive.ics.uci.edu/ml/