

The chordal graph polytope for learning decomposable models

Milan Studený and James Cussens

Institute of Information Theory and Automation of the CAS, Prague, Czech Republic
and Department of Computer Science, University of York, UK

(The 8-th) International Conference on *Probabilistic Graphical Models*
Lugano, Switzerland, September 8, 2016, 9:20-9:40

Summary of the talk

- 1 Introduction: motivation for learning decomposable models
- 2 Informal overview of results
- 3 Definitions: characteristic imset, chordal graph polytope
- 4 Example
- 5 Clutter inequalities and conjectures
- 6 Main result(s)
- 7 The separation problem in the cutting plane method
- 8 Conclusions

Motivation for learning decomposable models

Decomposable models are fundamental graphical models, which form the theoretical basis for the method of local computation. They are described by *chordal undirected graphs* and can be viewed as special cases of Bayesian network models, described by directed acyclic graphs.

There are various methods for learning the structure of decomposable models. This contribution deals with a *score-based approach*, where the task is to maximize some additively decomposable *score* (BIC or BDeu).

Specifically, we are interested in the *integer linear programming* (ILP) approach to structural learning (of decomposable models).

The idea behind this approach is to encode graphical models by certain vectors with integer components in such a way that the usual scores become affine/linear functions of the vector representatives.

Encoding decomposable models by characteristic imsets

There are more ways to encode Bayesian network models. The most successful one seems to be to encode them by *family-variable* vectors.



J. Cussens (2011). Bayesian network learning with cutting planes. *Uncertainty in Artificial Intelligence* 27, 153-160.

However, the approach discussed in this contribution is based on encoding the models by *characteristic imsets*, which are certain zero-one vectors with components indexed by subsets of the set of nodes N .



R. Hemmecke, S. Lindner, and M. Studený (2012). Characteristic imsets for learning Bayesian network structure. *International Journal of Approximate Reasoning* 53, 1336-1349.

This mode of representation leads to a particularly elegant way of encoding decomposable models.

Chordal graph polytope

Our approach leads to the study of the geometry of a polytope defined as the **convex hull of all characteristic imsets for chordal graphs**. This polytope has already been studied by Lindner in her thesis.



S. Lindner (2012). Discrete optimization in machine learning: learning Bayesian network structures and conditional independence implication. PhD thesis, TU Munich.

Following Lindner we name this polytope the “**chordal graph characteristic imset polytope**”, but abbreviate this to *chordal graph polytope*.

It is advantageous for the application of ILP maximization methods to have a **polyhedral description of the polytope** (= by means of linear inequalities), in other words, the characterization of its *facets*.

This theoretical contribution is devoted to an attempt to get a *complete facet description of the polytope*.

Clutter inequalities for the chordal graph polytope

We were able to compute the facets in cases of at most 5 nodes and to classify all facet-defining inequalities in those cases.

With the exception of a natural *lower bound inequality*, there is a one-to-one correspondence between the inequalities and the *clutters* of subsets of the node set N containing at least one singleton. Thus, we call these *clutter inequalities*.

We showed that every clutter inequality is indeed facet-defining for the chordal graph polytope. This establishes a sensible *conjecture* about the complete polyhedral description of the polytope.

Moreover, we offer a method to tackle an important *separation problem*: that is, given a non-integer solution to an LP relaxation problem, find a clutter inequality which (most) violates the current solution.

In the paper, we also discuss some preliminary empirical work, which only confirmed that to perform well-based computational experiments one has to solve a further theoretical task, namely, to design a method for *pruning the score* in case of decomposable models.

The concept of a characteristic imset

Each characteristic imset is an element of the vector space \mathbb{R}^Λ where $\Lambda := \{S \subseteq N : |S| \geq 2\}$.

Definition (characteristic imset)

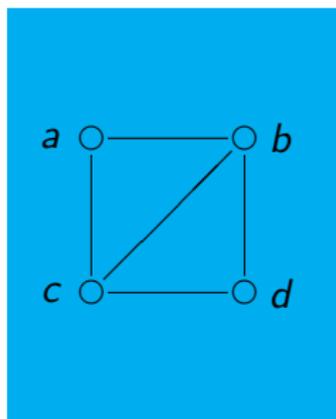
Given a chordal graph G over N , the *characteristic imset* of G is a zero-one vector c_G with components indexed by subsets S from Λ :

$$c_G(S) = \begin{cases} 1 & \text{if } S \text{ is a complete set in } G \text{ and } S \in \Lambda, \\ 0 & \text{for remaining } S \in \Lambda. \end{cases}$$

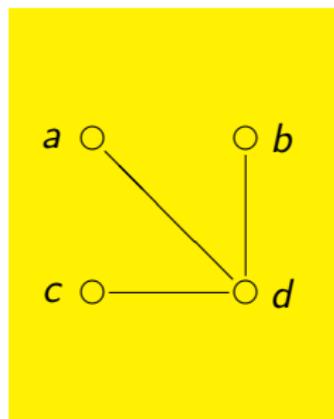
We adopt a convention that $c_G(L) = 1$ for any $L \subseteq N$ with $|L| = 1$.

Since decomposable models induced by chordal undirected graphs can be viewed as special cases of Bayesian network models each sensible scoring criterion is an affine function of the characteristic imset.

Example: characteristic imsets



$abcd$	0
abc	1
abd	0
acd	0
bcd	1
ab	1
ac	1
ad	0
bc	1
bd	1
cd	1
a	1
b	1
c	1
d	1



$abcd$	0
abc	0
abd	0
acd	0
bcd	0
ab	0
ac	0
ad	1
bc	0
bd	1
cd	1
a	1
b	1
c	1
d	1

Definition of the polytope

Definition (chordal graph polytope)

Let us introduce the *chordal graph polytope* over N as

$$D_N := \text{conv}(\{c_G : G \text{ chordal graph over } N\}),$$

where $\text{conv}(\cdot)$ is used to denote the convex hull.

Analogously, a *chordal graph polytope* with *cliques size limit* k , $2 \leq k \leq n = |N|$, can be introduced:

$$D_N^k := \text{conv}(\{c_G : G \text{ chordal graph over } N \text{ with clique size at most } k\}).$$

The dimension of D_N^k is $\sum_{\ell=2}^k \binom{n}{\ell}$. In particular, for the unrestricted polytope $D_N := D_N^n$ one has $\dim(D_N) = 2^n - n - 1$, while the most restricted polytope for learning *undirected forests* has $\dim(D_N^2) = \binom{n}{2}$.

Example: the case of 3 nodes

In the case $n = |N| = 3$, D_N has 8 vertices, namely 8 chordal graphs, and 8 facet-defining inequalities, decomposing into 4 permutation types.

With $N = \{a, b, c\}$, these are:

lower bound: $0 \leq c(\{a, b, c\})$,

2-to-3 monotonicity inequalities: $c(\{a, b, c\}) \leq c(\{a, b\})$,

upper bounds: $c(\{a, b\}) \leq 1$,

cluster inequality for 3-element set:

$$c(\{a, b\}) + c(\{a, c\}) + c(\{b, c\}) \leq 2 + c(\{a, b, c\}).$$

Note that the *cluster inequalities* (formulated in terms of family variables) have already occurred in the context of learning BN structure.



T. Jaakkola, D. Sontag, A. Globerson, and M. Meila (2010). Learning Bayesian network structure using LP relaxations. JMLR Workshop and Conference Proceedings 9, 358-365.

The cases of 4 and 5 nodes

In the case $n = |N| = 4$, the unrestricted polytope D_N has 61 vertices, that is, 61 chordal graphs. The number of facets is only 50, decomposing into 9 permutation types.

In the case $n = |N| = 5$, D_N has 822 vertices since there are 822 decomposable models. The number of its facets is again smaller, just 682, and they fall into 29 permutation types.

An interesting observation is this: in the case $n = |N| \leq 5$, with the exception of the lower bound $0 \leq c(N)$, all facet-defining inequalities for D_N have the form of a *generalized monotonicity*:

$$\sum_{S \subseteq N \setminus \{\gamma\}} \kappa(S) \cdot c(S \cup \{\gamma\}) \leq \sum_{S \subseteq N \setminus \{\gamma\}} \kappa(S) \cdot c(S)$$

where γ is a distinguished element of N and the $\kappa(S)$ are integer coefficients.

Introducing the clutter inequalities

A deeper fact is that the inequalities can be interpreted as inequalities induced by certain *clutters* of subsets of N , by which we mean classes of subsets of N that are inclusion incomparable.

Definition

Given a clutter \mathcal{L} of subsets of N which contains at least one singleton and satisfies $|\bigcup \mathcal{L}| \geq 2$, the corresponding **clutter inequality** for $c \in \mathbb{R}^{\mathcal{L}}$ has the form

$$1 \leq \sum_{\emptyset \neq B \subseteq \mathcal{L}} (-1)^{|B|+1} \cdot c(\bigcup B), \quad (1)$$

where a convention is applied that $c(L) = 1$ whenever $L \subseteq N$, $|L| = 1$.

We have re-written (1) in the form

$$1 \leq \sum_{S \subseteq N} \kappa_{\mathcal{L}}(S) \cdot c(S) \quad \text{and gave a formula for coefficients } \kappa_{\mathcal{L}}(*).$$

Example: what are the clutters in case of 3 nodes

With $N = \{a, b, c\}$, after the removal of the lower bound, there are 7 clutter inequalities decomposing into 3 types:

2-to-3 monotonicity inequalities: $c(\{a, b, c\}) \leq c(\{a, b\})$,

take $\mathcal{L} = \{ab, c\}$, (1) gives $1 \leq c(ab) + c(c) - c(abc)$
and, because of $c(c) = 1$, one gets $c(abc) \leq c(ab)$.

upper bounds: $c(\{a, b\}) \leq 1$,

take $\mathcal{L} = \{a, b\}$, (1) gives $1 \leq c(a) + c(b) - c(ab)$
and, since $c(a) = c(b) = 1$, one gets $c(ab) \leq 1$.

cluster inequality for 3-element set:

$$c(\{a, b\}) + c(\{a, c\}) + c(\{b, c\}) \leq 2 + c(\{a, b, c\}),$$

take $\mathcal{L} = \{a, b, c\}$, (1) gives

$$1 \leq c(a) + c(b) + c(c) - c(ab) - c(ac) - c(bc) + c(abc)$$

and one gets $c(ab) + c(ac) + c(bc) \leq 2 + c(abc)$.

The conjectures

We have the following conjecture, we know is valid in case $|N| \leq 5$.

Conjecture 1

For any $|N| \geq 2$, the set of facet-defining inequalities for $c \in D_N$ consists of the *lower bound* $0 \leq c(N)$ and the *clutter inequalities* (1) for such clutters \mathcal{L} of subsets of N containing at least one singleton and $|\bigcup \mathcal{L}| \geq 2$.

As concerns a prescribed clique size limit k , we conjecture the following.

Conjecture 2

For any $2 \leq k \leq n$, a polyhedral description of D_N^k is given by the lower bounds $0 \leq c(K)$ for $K \subseteq N$, $|K| = k$ and the inequalities (1) induced by clutters \mathcal{L} which are subsets of $\{L \subseteq N : |L| < k\}$, contain at least one singleton and satisfy $|\bigcup \mathcal{L}| \geq 2$.

Note that not every inequality from Conjecture 2 is facet-defining for D_N^k ; the problem of a precise characterization of facets of D_N^k is more subtle.

The main result of the paper

In the appendix of the paper we derive a formula for the LHS of the clutter inequality (1) in terms of the Möbius inversion of the characteristic imset. This allows to show easily the following result.

Corollary

Given a chordal graph G over N , $|N| \geq 2$, all inequalities from Conjecture 1 are valid for the characteristic imset c_G .

Nevertheless, we have a stronger theoretical result, namely that *every clutter inequality is facet-defining for D_N* , for any $|N| \geq 2$. However, its proof was omitted because of page limitation.

Idea of the cutting plane method

Since every clutter inequality is facet-defining for D_N , the number of inequalities describing D_N is super-exponential in $n = |N|$ and the use of a pure LP approach is not realistic.

Instead, *integer linear programming* (ILP) methods can be applied, specifically the *cutting plane method*. In this approach, the initial task is to solve an LP problem which is a *relaxation of the original problem*: namely to maximize the objective over a polyhedron P with $D_N \subseteq P$, where P is specified by a modest number of inequalities, typically, by some sub-collection of valid inequalities for D_N .

Moreover, facet-defining inequalities for D_N appear to be the most useful ones, leading to good overall performance.

Unless the optimal solution c^* to the relaxed problem has only integer components, one has to solve a *separation problem*, which is to find a linear constraint (a *cutting plane*) which separates c^* from D_N .

This new constraint is added and the method repeats starting from this new more tightly constrained problem.

Separation problem

The search limited to the *clutter inequalities* leads to the next task:

Given $c^ \notin D_N$ find clutter(s) \mathcal{L} such that the inequality (1) is (most) violated by c^* , in other words, we minimize*
$$\mathcal{L} \mapsto \sum_{S \subseteq N} \kappa_{\mathcal{L}}(S) \cdot c^*(S) \text{ over } \mathcal{L}.$$

Our idea is to re-formulate this in the form of a few auxiliary LP problems.

Specifically, if we fix a distinguished element $\gamma \in N$ and limit our search to clutters \mathcal{L} with $\{\gamma\} \in \mathcal{L}$ and $(\bigcup \mathcal{L}) \setminus \{\gamma\} \neq \emptyset$, then it leads to the task to solve an LP problem to minimize the above objective given by c^* over certain polytope.

We found a complete polyhedral description of that auxiliary polytope.

The details are omitted in this presentation; they can be found in the paper.

Conclusions

There are further supporting arguments for the conjectures. Specifically, we have derived from a classic matroid theory 1970 result by Edmonds that **a complete polyhedral description for D_N^2 consists of the lower bounds and the cluster inequalities**. Thus, Conjecture 2 is true in case $k = 2$.

We also have a promising ILP formulation for chordal graph learning using a subset of the facet-defining inequalities of D_N as constraints.

The big theoretical challenge remains: **to confirm/disprove Conjecture 1**. Even if confirmed, a further open problem is to characterize facet-defining inequalities for D_N^k , $2 \leq k \leq n$, within the clutter ones.

The preliminary empirical experiments indicate that **a further theoretical goal should be to develop special pruning methods** under the assumption that the optimal chordal graph is the learning goal. The subsequent goal, based on the result of pruning, can be to modify the proposed LP methods for solving the separation problem to become more efficient.

Some recent literature on learning DM

Two recent papers on ILP-based learning of decomposable models used a different binary coding of the models/graphs.

-  K. S. Sesh Kumar and F. Bach (2013). Convex relaxations for learning bounded-treewidth decomposable graphs. JMLR Workshop and Conference Proceedings 28 (1), 525-533.
-  A. Pérez, C. Blum, and J. A. Lozano (2014). Learning maximum weighted $(k + 1)$ -order decomposable graphs by integer linear programming. Lecture Notes in AI 8754, 396-408.

Moreover, two other recent papers devoted to learning decomposable models used encodings of junction trees.

-  J. Corander, T. Janhunen, J. Rintanen, H. Nyman, and J. Pensar (2013). Learning chordal Markov networks by constraint satisfaction. Advances in Neural Information Processing Systems 26, 1349-1357.
-  K. Kangas, T. Niinimäki, and M. Koivisto (2014). Learning chordal Markov networks by dynamic programming. Advances in Neural Information Processing Systems 27, 2357-2365.