Polyhedral Approach to Statistical Learning Graphical Models

Milan Studený
Department of Decision Making Theory
Institute of Information Theory and Automation of the ASCR
(ASCR = Academy of Sciences of the Czech Republic)
Pod Vodárenskou věží 4, Prague 18208, Czech Republic
studeny@utia.cas.cz

Raymond Hemmecke
Zentrum Mathematik
Technische Universität Munich
Boltzmannstrasse 3, Garching 85747, Germany
hemmecke@ma.tum.de

Jiří Vomlel
Department of Decision Making Theory
Institute of Information Theory and Automation of the ASCR
Pod Vodárenskou věží 4, Prague 18208, Czech Republic
vomlel@utia.cas.cz

Silvia Lindner
Department of Mathematics
University of Magdeburg
Universitätsplatz 2, Magdeburg 39106, Germany
lindner@mail.math.uni-magdeburg.de

The aim of the talk will be to explain how the statistical task to learn so-called Bayesian network structure from data leads to the study of a special polyhedron, and to report on what was found out about that polyhedron so far.

1. Motivation

Bayesian networks (BN) are basic graphical models, used widely both in statistics and artificial intelligence [6]. These statistical models of conditional independence structure are described by acyclic directed graphs whose nodes correspond to (random) variables in consideration.

The motivation for the research reported in this talk is *learning Bayesian network structure* from data by the method of maximization of a quality criterion. By a *quality criterion* is meant a real function $\mathcal{Q}$ of a BN structure (= of an acyclic directed graph G, usually) and of a database D. The value $\mathcal{Q}(G, D)$ "evaluates" how the BN structure determined by G fits the observed database D. A kind of standard example of such a criterion is Schwarz's *Bayesian information criterion* (BIC) [10], but there is also a bunch of "marginal likelihood" criteria that are motivated by a Bayesian viewpoint [7].

The basic idea of an algebraic and geometric approach to this topic, proposed in Chapter 8 of [11] and later developed in [13], 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 much higher than the number of (random) variables in consideration.

The point is that then every reasonable criterion $\mathcal{Q}$ for learning BN structures (namely, score equivalent [2] and decomposable [3] one) is an affine function of the standard imset. More specifically, one has

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

where $s_D^{\mathcal{Q}}$ is a real number, $t_D^{\mathcal{Q}}$ a vector of the same dimension as the standard imset u_G (these parameters both depend solely on the database D and the criterion $\mathcal{Q}$) and $\langle *, * \rangle$ denotes the scalar product. 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 [12].

2. Standard imset polytope

The main result of [13] is that the set of standard imsets over a fixed set of (random) variables N is the set of vertices (= extreme points) of a certain polytope P, called the *standard imset polytope* in the sequel. Thus, as every reasonable quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on the respective Euclidean space, the task to maximize $\mathcal{Q}$ over BN structures is equivalent to the task to maximize an affine function over the above-mentioned polytope P.

This maximization problem has been treated thoroughly within the linear programming community. The intention to apply linear programming methods in the area of learning BN structures motivated several open mathematical questions concerning that polytope (see the conclusions of [13]). Here we report on some of them.

3. Towards the outer description of the polytope

A standard tool to solve linear programming problems is the *simplex method* [8]. In order to apply the (classic) simplex method, one needs an explicit *outer description* of the polytope via finitely many linear inequalities (= 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 $|N|$, one definitely needs to classify these inequalities implicitly.

One of our research directions was to (try to) classify necessary linear inequality constraints on P. In [14] we analyzed the case $|N| = 4$ and came with the classification of inequalities into three classes, namely:

- trivial equality constraints

- inequality constraints that correspond to (standardized) *extreme supermodular functions* on the power set of N,
- inequality constraints which correspond to classes of subsets of N that are closed under supersets.

We conjecture that these constraints already characterize the polytope P and the current task is either to confirm or disprove this conjecture for $|N| = 5$.

4. 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 [13] the concept of the *geometric neighborhood* for standard imsets, and, consequently, for BN structures.

More specifically, two standard imsets are called *geometric neighbors* if the line segment connecting them is 2-dimensional face (= an 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 have succeeded to compute the geometric neighborhood for $|N| = 3, 4, 5$. As a by-product we compared it for $|N| = 3$ with the *inclusion neighborhood* [7], which is at the core of current computer science search techniques (for maximization of a quality criterion), like so-called GES algorithm [3]. 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 a quality criterion. Actually, we think that this is an inevitable defect of the inclusion neighborhood, which may occur whenever a special statistical *data faithfulness assumption* is not guaranteed.

The result of our analysis of the geometric neighborhood in the case $|N| = 4$ is an electronic catalogue of types of geometric neighbors [15]. The catalogue is 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.

5. Lattice points in the polytope

Raymond Hemmecke was interested in the question of how “thick” the standard imset polytope P is. More specifically, he made some computations to find out whether there exists a lattice point in its interior for $|N| \leq 5$ and the result was negative.

This led him to a conjecture that every lattice point in the standard imset polytope is already a standard imset. In [14] we confirmed his conjecture. The final proof is surprisingly simple. The idea is to apply certain one-to-one affine 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 0-1 hypercube, the above statement is immediate.

6. Characteristic imset

The image of the standard imset u_G by that transformation, called the *characteristic imset* (for G) in [5], may occur to be even better algebraic BN structure representative. It is much closer to the graphical description because some of its components directly correspond to adjacencies in G and allows to identify so-called *immoralities* in G, which together with the adjacencies characterize the BN structure graphically [1].

One of our further research plans is to apply the geometric approach to restricted learning graphical models. If we restrict our attention to so-called *decomposable models* [6] then the characteristic imset has quite simple form [5].

The situation is particularly transparent in the case of undirected forests, when adjacencies in the graph correspond to ones in the characteristic imset. The corresponding (transformed) polytope has already been studied in matroid theory and both its outer description and edges were characterized [9]. This allows us to give the geometric interpretation to the classic heuristic procedure for learning undirected trees [4].

References

[1] S. A. Andersson, D. Madigan, M. D. Perlman, A characterization of Markov equivalence classes for acyclic digraphs, The Annals of Statistics 25 (1997) 505-541.

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

[3] D. M. Chickering, Optimal structure identification with greedy search, Journal of Machine Learning Research 3 (2002) 507-554.

[4] C. K. Chow, C. N. Liu, Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory 14 (1968) 462-467.

[5] R. Hemmecke, S. Lindner, M. Studený, J. Vomlel, Characteristic imsets for learning Bayes network structures, in preparation.

[6] S. L. Lauritzen, Graphical Models, Clarendon Press, 1996.

[7] R. E. Neapolitan, Learning Bayesian Networks, Pearson Prentice Hall, 2004.

[8] A. Schrijver, Theory of Linear and Integer Programming, John Wiley, 1986.

[9] A. Schrijver, Combinatorial Optimization - Polyhedra and Efficiency, volume B, Springer Verlag, 2003.

[10] G. Schwarz, Estimation the dimension of a model, The Annals of Statistics 6 (1978) 461-464.

[11] M. Studený, Probabilistic Conditional Independence Structures, Springer Verlag, 2005.

[12] M. Studený, Mathematical aspects of learning Bayesian networks: Bayesian quality criteria, research report n. 2234, Institute of Information Theory and Automation, Prague, December 2008.

[13] M. Studený, J. Vomlel, R. Hemmecke, A geometric view on learning Bayesian network structures, forthcoming, International Journal of Approximate Reasoning (2010), `doi:10.1016/j.ijar.2010.01.014`.

[14] M. Studený, J. Vomlel, On open questions in the geometric approach to structural learning Bayesian nets, submitted to International Journal of Approximate Reasoning, special issue WUPES'09.

[15] J. Vomlel, M. Studený, The catalogue of types of geometric neighbors over four variables, available at `http://staff.utia.cas.cz/vomlel/imset/catalogue-diff-imsets-4v.html`.