"Seventh Valencia International Meeting on Bayesian Statistics", June 2-6, 2002, Tenerife (Canary Islands, Spain)

An algebraic approach to learning Bayesian networks

Milan Studený
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Prague

Graphical models of *conditional independence* (CI) structures are popular both in the area of artifical intelligence (AI) and in statistics.

This contribution concerns the problem of learning graphical models described by acyclic directed graphs, known as *Bayesian networks* in AI. Some of Bayesian approaches to learning Bayesian networks from data are based on the idea of maximalization of a *score metric* and some of them use the method of *local search* for its maximalization.

In this contribution an *algebraic point of view* on this method is presented. The main idea is as follows: every Bayesian network model can be represented by a certain integral vector, named *standard structural imset* and the value of a (reasonable) score metric is then a linear function of this vector and of another real vector depending on data only.

Moreover, (reasonable) moves to neighbouring (Bayesian network) models used in the method of local search can also be represented by very simple integral vectors, named *elementary imsets*, which correspond to elementary CI statements.

Bayesian networks I

Every Bayesian network can be viewed as a specific statistical model, that is a class of probability distributions.

G acyclic directed graph having N as the set of nodes ($N \neq \emptyset$ finite)

Commonly used (but grammatically misleading) phrase **directed acyclic graph** *leads to generally accepted abbreviation* **DAG**.

Supposing $\mathbf{X}_i, i \in N$ are respective sample (measurable) spaces the respective class $\mathcal{M}(G)$ of probability distributions on the joint sample space $\mathbf{X}_N$ can be introduced in two (typically) equivalent ways.

A Using recursive factorization formula

- $\forall\, A \subseteq N \quad \mathbf{X}_A = \prod_{i \in A} \mathbf{X}_i$ the *sample space for* A
- Given $x_N \in \mathbf{X}_N$ the symbol x_A denotes the *projection* of x_N onto $\mathbf{X}_A$
- Let f denote the density of a probability distribution P on $\mathbf{X}_N$ with respect to a suitable dominating product measure. According to a common convention the *marginal* and *conditional densities* will be denoted by the same symbol and their arguments will indicate what is actually meant.
  For example, if $A, B \subseteq N$ are disjoint then $f(x_A|x_B)$ is used to denote the value of the conditional density on $\mathbf{X}_A$ given x_B (= the configuration $x_B \in \mathbf{X}_B$ of values for B) in the configuration $x_A \in \mathbf{X}_A$.
- $\forall\, i \in N \quad pa\,(i) = \{b \in N\,;\, b \to i \text{ in } G\}$ *parents* of a node i (in G)
  If it necessary to indicate the graph one writes $pa_{\,G}(i)$.

$$\mathcal{M}(G) = \{\, P \text{ on } \mathbf{X}_N\,;\, \forall\, x_N \in \mathbf{X}_N \;\; f(x_N) = \prod_{i \in N} f(x_i|x_{pa\,(i)}) \,\}\,.$$

In case of finite $\mathbf{X}_i$'s the values of conditional densities serve as traditional parameters for parametric description of $\mathcal{M}(G)$.

Bayesian networks II

B Models of conditional independence structure

- $\mathcal{T}(N) = \{\, \langle A, B|C\rangle \,;\, A, B, C \subseteq N \text{ pairwise disjoint} \,\}$
  the class of (disjoint) *triplets* over N
- $A \perp\!\!\!\perp B \,|\, C \;[P]$ denotes respective *CI statement* relative to P, i.e. the statement saying that for a random vector $[\xi_i]_{i\in N}$ on $\mathbf{X}_N$ having distribution P its subvectors $[\xi_i]_{i\in A}$ and $[\xi_i]_{i\in B}$ are conditionally independent given $[\xi_i]_{i\in C}$.

Graphical separation criteria There are two traditional equivalent criteria which ascribe to a DAG G the respective formal independence model. One of them is the *moralization criterion* [Lauritzen et.al. 1990], the other one is the *d-separation criterion* [Pearl 1988] (here $d-$ means 'directional').

- A node c is a *descendant* of a node i in G if there exists a directed path $i = d_1 \to d_2 \ldots \to d_n = c$, $n \geq 1$ in G from i to c (possibly $i = c$).
- Given $\langle A, B|C\rangle \in \mathcal{T}(N)$ one says that C *d-separates* between A and B in G if for every path in G (not necesarily a directed path) from a node in A to a node in B there exists a node i on the path such that
  - either i is a *collider* on the path: $a \to i \leftarrow b$ and none of its descendants belongs to C
  - or i is not a collider on the path and belongs to C.
- *Formal independence model*

  $\mathcal{I}(G) = \{\langle A, B|C\rangle \in \mathcal{T}(N)\,;\, C \text{ } d\text{-separates between } A \text{ and } B \text{ in } G\,\}\,.$

$$\mathcal{M}(G) = \{\, P \text{ on } \mathbf{X}_N \,;\, A \perp\!\!\!\perp B \,|\, C \;[P] \text{ if } \langle A, B|C\rangle \in \mathcal{I}(G)\,\}\,.$$

Score metrics

One of the approaches to learning Bayesian networks from data (in the form of a sequence of elements of $\mathbf{X}_N$ assumed to be a 'realization' of an i.i.d. sequence of random variables with a shared distribution P) uses the idea of *score metric*.

However, there are other approaches, for example approaches based on statistical tests of CI statements.

The idea is that a statistician chooses a function S which ascribes a real number $S(G, D)$ to every DAG $G \in \mathsf{DAGS}(N)$ (= the collection of acyclic directed graphs having N as the set of nodes) and every data $D \in \mathsf{DATA}(N, d)$ (= the collection of all sequences of elements of $\mathbf{X}_N$ of length $d \geq 1$) which measures how suitable is the statistical model determined by G for explanation of the occurence of data D, shortly how the DAG G fits the data D. This function is here named the *score metric*.

The basic aim is clear: the higher the value of $S(G, D)$ is the better the statistical model determined by G should fit the data D. However, some usual score metrics have also *penalization terms* which somehow (negatively) reflect the complexity of a model (measured for example by the 'number' of free parameters). Therefore, from a mathematical point of view the resulting task to maximize $S(G, D)$ over G's for fixed D.

Equivalence of DAGs

Since the aim is to find 'the best' statistical model $\mathcal{M}(G)$ it is not reasonable to distinguish between *equivalent DAGs*, that is DAGs K and L for which $\mathcal{M}(K) = \mathcal{M}(L)$.

Note that if the sample spaces are non-trivial (and the distribution framework is sufficiently wide) then this is characterized by the condition that the induced formal independence models coincide: $\mathcal{I}(K) = \mathcal{I}(L)$.

The situation $\mathcal{I}(K) = \mathcal{I}(L)$ can be characterized directly in graphical terms as well.

Therefore, a natural assumption is that the metric S is *score-equivalent* which means $S(K, D) = S(L, D)$ for every data $D \in \mathsf{DATA}(N, d)$ and $K, L \in \mathsf{DAGS}(N)$ such that $\mathcal{I}(K) = \mathcal{I}(L)$.

Decomposable criteria

Further reasonable assumption is that the score metric factorizes according to the DAG in a way which is analogous to the (recursive) factorization formula - see [Chickering 2002].

- $\forall\, A \subseteq N \;\; D \in \mathsf{DATA}\,(N, d)$ $\quad D_A$ denotes the *projection* of D onto $\mathbf{X}_A$

$$x^1, \ldots, x^d \quad \mapsto \quad x^1_A, \ldots, x^d_A$$

A score metric S is called *decomposable* if there exists a collection of real functions $s_{i|B}$ on $\mathsf{DATA}\,(\{i\} \cup B, d)$ where $i \in N$ and $B \subseteq N \setminus \{i\}$ such that $\forall\, G \in \mathsf{DAGS}\,(N) \;\; D \in \mathsf{DATA}\,(N, d)$

$$S(G, D) = \sum_{i \in N} s_{i|pa\,(i)}\left(D_{\{i\} \cup pa\,(i)}\right).$$

I myself advocate for the following concept. A score metric is *regular* if there exists a collection of real functions t_A on $\mathsf{DATA}\,(A, d)$ where $A \subseteq N$ such that
$\forall\, G \in \mathsf{DAGS}\,(N) \;\; D \in \mathsf{DATA}\,(N, d)$

$$S(G, D) = \sum_{i \in N} t_{\{i\} \cup pa\,(i)}(D_{\{i\} \cup pa\,(i)}) \;-\; t_{pa\,(i)}(D_{pa\,(i)}).$$

Observation 1 A score metric is regular iff it is score-equivalent and decomposable.

Examples of score metrics I

Typical method of derivation of a suitable score metric is the method of maximized likelihood [Cowell et.al. 1999]. This leads to *maximum log-likelihood* criterion which is the maximum of the logarithm of the likelihood function $l(D,P)$ (= the probability of occurence of data D provided that P 'generates' the data) over all probability distributions $P \in \mathcal{M}(G)$ in the model:

$$\mathsf{MLL}\,(G,D) = \max\{\,\ln\, l(D,P)\,;\; P \in \mathcal{M}(G)\,\}\,.$$

If one has discrete data (i.e. finite $\mathbf{X}_i$'s) and the order of items in data is important (that is one distinguishes between data $x^1 = y$, $x^2 = z$ and $x^1 = z$, $x^2 = y$) then a direct formula can be written.

Conventions

i code of a node $i \in N$

k code of a value of a variable $i \in N \quad k = 1, \ldots, r(i)$, $r(i) = |\mathbf{X}_i|$

j code of a value of a parent configuration for $i \in N \quad j = 1, \ldots, q(i,G)$, $q(i,G) = |\mathbf{X}_{pa\,(i)}|$

Then

$$\mathsf{MLL}\,(G,D) = \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot \ln\frac{d_{ijk}}{d_{ij}}\,,$$

where d_{ijk} respectively d_{ij} denotes the number of occurences of the respective configuration in data $D \in \mathsf{DATA}\,(N,d)$.

Penalized derived criterion is *Akaike's Information Criterion*:

$$\mathsf{AIC}\,(G,D) = \mathsf{MLL}\,(G,D) - d(G)\,,$$

where $d(G)$ denotes the number of free parameters in the statistical model $\mathcal{M}(G)$ given by

$$d(G) = \sum_{i\in N}(\,r(i)-1\,)\cdot\sum_{\ell\in pa\,(i)} r(\ell)\,.$$

Examples of score metrics II

Another popular criterion is Jeffreys-Schwarz criterion, sometimes called *Bayesian Information Criterion*:

$$\mathsf{BIC}\,(G, D) = \mathsf{MLL}\,(G, D) - \frac{1}{2} \cdot d(G) \cdot \ln d\,.$$

One can easily show that all these three criteria are regular.

In Bayesian approach one specifies a prior density π_G on the set of parameters for each statistical model $\mathcal{M}(G)$ into consideration. One usually puts additional assumptions on these priors and considers convenient forms for them - see e.g. [Spiegelhalter Lauritzen 1990]. Moreover, the priors for distinct models are implicitly supposed to be related.

Consequently, having fixed a collection of priors by integrating the likelihood function according to them (and taking possibly the logarithm of the result) other reasonable score metrics can be derived. For example, the *logarithm of the marginal likelihood*:

$$\mathsf{LML}\,(G, D) = \ln \int_{\mathcal{M}(G)} l(P, D)\; d\,\pi_G(P)$$

can be shown to be a regular criterion on suitable asumptions, I guess.

The method of local search

It may be computationally demanding task to evaluate the score metric in general. To avoid these problems the method of *local search* was proposed [Meek 1997], [Chickering 2002].

The idea is to introduce a suitable concept of *neighbourhood* for statistical models into consideration. Thus, the class of Bayesian network models over N can be understood as a (huge) state space whose states are equivalence classes of DAGs. Therefore, the *search space* is the class $\{ \mathcal{M}(G) \, ; \, G \in \mathsf{DAGS}\,(N) \}$ endowed with a suitable symmetric binary relation 'being a neighbour'.

One starts the procedure in a specific state and in every step computes the change in the value of the score metric only for some of (a limited number of) its neighbouring states. For reasonable score metrics (= regular ones) the change in score is easy to compute since it depends only on a few 'local' terms. Then one 'moves' to the state with the highest increase in the score.

This method certainly leads to a 'local maximum' in general but on some (relatively strong) assumptions it is guaranteed to find the right (= 'generating') model [Meek 1997].

To ensure the convergence of the above procedure the neighbourhood of a state in the chosen search space has to involve certain minimal neighbourhood (which is justified from a theoretical point of view).

Inclusion of DAGs

Natural idea of defining theoretically justified concept of neighbourhood is based on the inclusion of statistical models.

Let K and L are DAGs over N. If mild assumptions on the sample spaces (and the distribution framework) are fulfilled then the inclusion $\mathcal{M}(L) \subseteq \mathcal{M}(K)$ is equivalent to the condition $\mathcal{I}(K) \subseteq \mathcal{I}(L)$.

One says that two DAGs K and L over N are *neighbours* in sense of *inclusion boundary* and uses notation $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ if

- $\mathcal{I}(K) \subset \mathcal{I}(L)$ (i.e. $\mathcal{I}(K) \subseteq \mathcal{I}(L)$ but $\mathcal{I}(K) \neq \mathcal{I}(L)$),
- there is no DAG G over N such that $\mathcal{I}(K) \subset \mathcal{I}(G) \subset \mathcal{I}(L)$.

Well, it is the matter of taste whether one says then that L is an upper neighbour of K and K the lower neighbour of L or conversely.

It seems to be quite difficult task to characterize inclusion of DAGs in graphical terms. Note that a conjecture about graphical characterization of inclusion (boundary) [Meek 1997] was recently confirmed [Chickering 2002].

Internal computer representation

Traditional methods of internal computer representation of equivalence classes of DAGs use partially directed graphs (i.e. graphs having some edges directed and some undirected ones):

- *patterns* [Meek 1997] may have semi-directed cycles
- *essential graphs* [Andersson et.al. 1977], sometimes named 'completed patterns', are acyclic.

In this contribution a non-graphical method of computer representation of equivalence classes of DAGs, in particular a method of computer representation of Bayesian network statistical models, is proposed.

The main idea is to represent them by certain integral vectors, named *structural imsets*. This can be viewed as an example of application of a more general method of description of probabilistic CI structures by integral vectors of this kind [Studený 2001]. The actual proposal is to represent every equivalence class of DAGs by unique *standard structural imset*.

Structural imsets

By an *imset* over N is understood an integral function on $\mathcal{P}(N)$, the power set of N.

The word 'imset' is an abbreviation for **i***nteger-valued* **m***ulti***set**.

- $\forall A \subseteq N \quad \delta_A$ denotes the *identificator* of A

$$\delta_A(A) = 1, \quad \delta_A(B) = 0 \ \text{ for } B \subseteq N,\ B \neq A\,.$$

- $\mathcal{T}_\epsilon(N) = \{\langle a, b|C\rangle\,;\ a, b \in N,\ a \neq b,\ C \subseteq N \setminus \{a, b\}\,\}$ the class of *elementary triplets* over N

These triplets correspond to elementary CI statements from a certain theoretical point of view.

Given $\langle a, b|C\rangle \in \mathcal{T}_\epsilon(N)$ the respective *elementary imset* $u_{\langle a,b|C\rangle}$ is given by the formula

$$u_{\langle a,b|C\rangle} = \delta_{\{a,b\}\cup C} + \delta_C - \delta_{\{a\}\cup C} - \delta_{\{b\}\cup C}\,.$$

The class of elementary imsets will be denoted by $\mathcal{E}(N)$. An imset is called *combinatorial* if it is a combination of elementary imsets with non-negative integral coefficients:

$$u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v \quad \text{where } k_v \in \mathbb{Z}^+\,.$$

An imset w is a *structural imset* if for some natural number $n \in \mathbb{N}$ the multiple $n \cdot w$ is a combinatorial imset.

There is a polynomial algorithm of testing whether an imset is a combinatorial imset. On the other hand, testing structural imsets could appear to be a problem unless the following conjecture is confirmed. It is an open question whether every structural imset is a combinatorial imset. The conjecture was verified in case $|N| \leq 4$.

Standard structural imsets

Given a DAG $G \in \mathsf{DAGS}(N)$ the respective *standard structural imset* u_G is given by the formula

$$u_G = \delta_N - \delta_\emptyset + \sum_{i \in N} \delta_{pa(i)} - \delta_{\{i\} \cup pa(i)} .$$

Actually, it is a combinatorial imset.

The concept of standard structural imset offers a simple non-graphical method of testing equivalence of DAGs.

Theorem 1 Two DAGs K and L over N are equivalent iff $u_K = u_L$.

In particular, standard structural imsets can serve as representatives of equivalence classes of DAGs. However, this concept is also useful for the following fact.

Observation 2 Let S be a regular score metric. Then there exists a real function $t : \mathcal{P}(N) \times \mathsf{DATA}(N, d) \mapsto \mathbb{R}$ such that $\forall\, G \in \mathsf{DAGS}(N)\ D \in \mathsf{DATA}(N, d)$

$$S(G, D) = t(N, D) - t(\emptyset, D) - \sum_{A \subseteq N} t(A, D) \cdot u_G(A)$$
$$\equiv constant(D) - \langle t_D, u_G \rangle .$$

Actually, the value of $t(A, D) \equiv t_D(A)$ *depends on* D *only through* D_A.

Thus, the value of a score metric is a linear function of u_G, more precisely it is a constant depending on data plus the scalar product of the vector u_G and of another real vector t_D which represents data in this framework.

Change in score

Morevover, standard structural imsets allow one to characterize inclusion of DAGs using an algebraic relation.

Observation 3 Let K and L are DAGs over N. Then $\mathcal{I}(K) \subseteq \mathcal{I}(L)$ iff $u_L - u_K$ is a combinatorial imset.

Note that this result can be easily derived as a consequence of the validity of above mentioned Meek's conjecture about graphical characterization of inclusion. Recall that there exists a polynomial algorithm for testing combinatorial imsets.

One can say even more, namely that the neighbourhood in sense of inclusion boundary can be characterized using an algebraic relation. Finally, the change in the value of a score metric is also a linear function of a certain elementary imset.

Observation 4 Let S be a regular score metric and K, L are DAGs over N. Then $\mathcal{I}(K) \sqsubset \mathcal{I}(L)$ iff $u_L - u_K$ is an elementary imset. Of course, this elementary imset $u_{\langle a,b|C\rangle}$ is uniquely determined by K and L and $\forall\, D \in \mathsf{DATA}\,(N, d)$

$$S(K, D) - S(L, D) = \sum_{A \subseteq N} t(A, D) \cdot u_{\langle a,b|C\rangle}(A) \equiv \langle t_D, u_{\langle a,b|C\rangle}\rangle$$

where $t(A, D)$ is the function mentioned in Observation 2.

Since $u_{\langle a,b|C\rangle}$ has only 4 non-zero values one has

$$\langle t_D, u_{\langle a,b|C\rangle}\rangle = t_D(\{a, b\} \cup C) + t_D(C) - t_D(\{a\} \cup C) - t_D(\{b\} \cup C)\,.$$

Conclusions

The presented approach allow one

- to represent every Bayesian network model $\mathcal{M}(G)$ by a certain integral vector u_G called *standard structural imset*,
- to represent respective data by a real vector t_D,
- express the value of a resonable score metric (namely a *regular* one) as the scalar product of these two vectors (plus a constant),
- interpret the change in score between neighbouring models in terms of an elementary CI statement and express it as a scalar product as well.

Moreover, the approach allows one to characterize inclusion of DAGs in an algebraic way and brings a clear 'algebraic' point of view on the problem.

It is my hope that this point of view will lead to an alternative method of maximalization of score metric in future.

Finally, the approach can be viewed as a specific application of a more general method of mathematical description of probabilistic CI structures by means of structural imsets [Studený 2001]. Thus, perhaps this algebraic approach can be extended to learning other classes of graphical models.

But the present problem (of time) is that:

- respective full paper on this topic is to be written ...
- many interesting open questions need to be answered ...

Literature

- S.A. Andersson, D. Madigan, M.D. Perlman: A characterization of Markov equivalence classes for acyclic digraps, *Annals of Statistics* 25 (1997) 505-541.
- D.M. Chickering: Optimal structure identification with greedy search, submitted to *Jornal of Machine Learning Research* (2002).
- R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter: Probabilistic Networks and Expert Systems, Springer 1999.
- S.L. Lauritzen, A.P. Dawid, B.N. Larsen, H.-G. Leimer: Independence properties of directed Markov fields, *Networks* 20 (1990) 491-505.
- C. Meek: Graphical models, selecting causal and statistical models, PhD thesis, Carnegie Mellon University 1997.
- J. Pearl: Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference, Morgan Kaufmann 1988.
- D.J. Spiegelhalter, S.L. Lauritzen: Sequential updating of conditional probabilities on directed acyclic structures, *Networks* 20 (1990) 579-606.
- M. Studený: On mathematical description of probabilistic conditional independence structures, thesis for DrSc degree, Institute of Information Theory and Automation, Prague 2001.