On structural imsets for describing and learning graphical models

Milan Studený

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

Pacific Causal Inference Conference 2022, Beijing, China
online session *Imsets and Applications to Graphical Models*
(organized by Robin Evans)

Summary of the talk

Motivation: graphical models of CI structure

Conditional independence (CI) concept is at the core of graphical statistical models. Besides their basic *CI interpretation*, graphical models typically admit an extended *causal interpretation*.

J. Pearl (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo.

S. L. Lauritzen (1996). Graphical Models. Clarendon Press, Oxford.

As n increases, graphs over n nodes cannot faithfully describe all possible CI structures induced by n discrete random variables.

This was the motivation to develop a *linear-algebraic approach* to describe CI structures using certain vectors whose components are integers.

M. Studený (2005). Probabilistic Conditional Independence Structures. Springer, London.

Such vectors can also be used to describe the graphical models.

Introduction: structural imsets

These vectors are called *structural imsets*.

The method offers a sufficient condition for verifying the probabilistic implication between CI statements – the respective inference mechanism is based on **linear algebraic operations** with imsets.

The imsets can also be utilized in context of *learning graphical models*: they can be used as (unique) linear-algebraic representatives of (Markov) equivalence classes of graphs.

Also, there is a (deep theoretical) connection of structural imsets to the class of *supermodular set functions* on subsets of the variable set. The supermodular/submodular set functions play a crucial role in optimization.

S. Fujishige (2005). Submodular Functions and Optimization (2nd edition). Elsevier, Amsterdam.

Conditional independence statements/models

- N ... a finite set of variables
- $\mathcal{P}(N) := \{A : A \subseteq N\}$ ... the power set of N
- AB ... instead of $A \cup B$ for $A, B \subseteq N$
- i ... will be a shorthand for $\{i\}$ if $i \in N$

Definition (conditional independence)

Let $\langle A, B|C\rangle$ be a triplet of pairwise disjoint subsets of N. Let P be a discrete probability distribution over N. We say that A is *conditionally independent* of B given C with respect to P and write $A \perp\!\!\!\perp B \,|\, C \; [P]$ if

$$P(\mathbf{a}|\mathbf{bc}) = P(\mathbf{a}|\mathbf{c}) \quad \text{for configurations } \mathbf{a}, \mathbf{b}, \mathbf{c} \text{ for } A, B, C \text{ with } P(\mathbf{bc}) > 0\,.$$

A statement of this form will be called a *CI statement*.
The (discrete probabilistic) *CI model* induced by P is as follows:

$$\mathcal{M}_P := \{\, \langle A, B|C\rangle : \quad A \perp\!\!\!\perp B \,|\, C \; [P] \,\}\,.$$

Conditional independence in terms of densities

There are various equivalent definitions of CI which can be extended beyond the framework of discrete distributions.

Consider a *marginally continuous* probability distribution P on $\mathrm{X}_N := \prod_{i\in N} \mathrm{X}_i$: $P \ll \mu$, where μ is a dominating product measure on X_N (Lebesgue measure on $\mathbb{R}^N$ or the counting measure on X_N). For any $A \subseteq N$, the marginal P_A is described by its *marginal density* $f_A : \mathrm{X}_N \to [0, \infty)$, depending on A-factors only.

Lemma (CI in terms of marginal densities/factorization)

Given a marginally continuous distribution P one has $A \perp\!\!\!\perp B \,|\, C$ $[P]$ iff

$$f_{ABC}(x) \cdot f_C(x) = f_{AC}(x) \cdot f_{BC}(x) \quad \text{for } \mu\text{-a.e. } x \in \mathrm{X}_N\,.$$

Another condition is that there are functions $g_{AC} : \mathrm{X}_N \to \mathbb{R}$ depending on AC-factors and $h_{BC} : \mathrm{X}_N \to \mathbb{R}$ depending on BC-factors such that

$$f_{ABC}(x) = g_{AC}(x) \cdot h_{BC}(x) \quad \text{for } \mu\text{-a.e. } x \in \mathrm{X}_N\,.$$

The concept of an imset

Definition (imset)

An *imset* u (over N) is a function $u : \mathcal{P}(N) \mapsto \mathbb{Z}$.

imset = an abbreviation for **i** nteger-valued **m** ulti **set**

We will regard an imset u over N as a vector whose components are integers and are indexed by subsets of N: $u \in \mathbb{Z}^{\mathcal{P}(N)}$. Any real set function $m : \mathcal{P}(N) \to \mathbb{R}$ will be interpreted as a (real) vector in the same way: $m \in \mathbb{R}^{\mathcal{P}(N)}$. The symbol $\langle m, u \rangle$ will then denote the scalar product of two vectors of this type:

$$\langle m, u \rangle := \sum_{A \subseteq N} m(A) \cdot u(A) \,.$$

Given $A \subseteq N$, the symbol δ_A will denote a special imset given by:

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

Elementary and semi-elementary imsets

Definition (translation of a CI statement, elementary imset)

Given $A \perp\!\!\!\perp B \,|\, C$, the corresponding imset representing this CI statement will be a *semi-elementary imset*

$$u_{\langle A,B|C\rangle} := \delta_{ABC} + \delta_C - \delta_{AC} - \delta_{BC} \,.$$

By an *elementary imset* is meant an imset of the form

$$u_{\langle a,b|C\rangle} := \delta_{abC} + \delta_C - \delta_{aC} - \delta_{bC} \,,$$

where $C \subseteq N$ and $a, b \in N \setminus C$ are distinct.

The class of elementary imsets over N will be denoted by $\mathcal{E}(N)$.

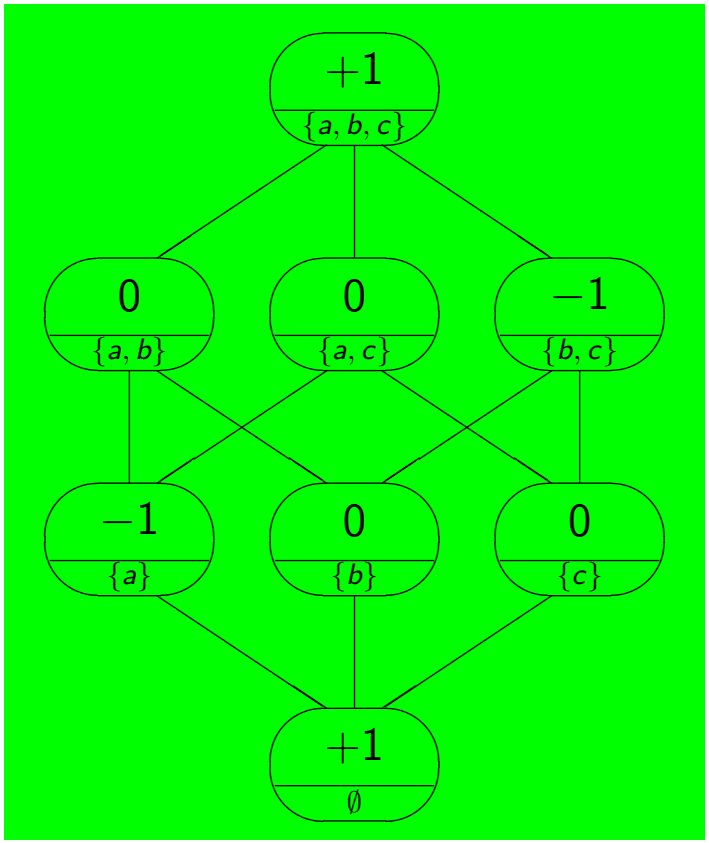
Every semi-elementary imset is a combination of elementary imsets with non-negative integers as coefficients.

Example: visualization of imsets

If the number of variables is low, an imset can be visualized in the form of an enriched Hasse diagram.

$N = \{a, b, c\}$

$u = u_{\langle a,b|c\rangle} + u_{\langle a,c|\emptyset\rangle} \equiv u_{\langle a,bc|\emptyset\rangle}$ it is a semi-elementary imset



Combinatorial and structural imsets

Let us call an imset *combinatorial* if it is a combination of elementary imsets with non-negative integers as coefficients.

Definition (structural imset)

A *structural imset* is an imset u which is a combination of elementary imsets with non-negative rational coefficients. Equivalently, one of its multiples by a positive natural number is a combinatorial imset:

$$n \cdot u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v \quad \text{for some } n \in \mathbb{N} \text{ and } k_v \in \mathbb{Z}^+ .$$

The class of structural imsets over N will be denoted by $\mathcal{S}(N)$.

Every combinatorial imset is structural but the converse is not true:

R. Hemmecke, J. Morton, A. Shiu, B. Sturmfels, O. Wienand (2008). Three counter-examples on semi-graphoids. Combinatorics, Probability and Computing 17, 239-257.

Markov condition relative to a structural imset

An analogue of graphical "separation" criteria is a special **linear-algebraic criterion** for a structural imset over N (replacing a graph over N).

Definition (representation of a CI statement within a imset)

Given a structural imset u over N and a triplet $\langle A, B|C\rangle$ of pairwise disjoint subsets of N we will write $A \perp\!\!\!\perp B \mid C \ [u]$ if
there exists $k \in \mathbb{N}$ such that $k \cdot u - u_{\langle A,B|C\rangle}$ is a structural imset.

Definition (Markovian distribution)

Given a structural imset u over N, the induced CI model is as follows:

$$\mathcal{M}_u := \{ \langle A, B|C\rangle : \quad A \perp\!\!\!\perp B \mid C \ [u] \}.$$

A probability distribution P is *Markovian* with respect to u if $\mathcal{M}_u \subseteq \mathcal{M}_P$.
If $\mathcal{M}_u = \mathcal{M}_P$ then P is called *perfectly Markovian* with respect to u.

Markov property interpreted as a factorization formula

The *multiinformation* of a probability distribution P over N is the Kullback-Leibler divergence $H(P \,\|\, \prod_{i\in N} P_i)$. The multiinformation function $m_P \in \mathbb{R}^{\mathcal{P}(N)}$ ascribes the multiinformation of the marginal P_A to any set $\emptyset \neq A \subseteq N$; $m_P(\emptyset) := 0$.

A distribution with finite multiinformation is marginally continuous.

Theorem (Markovness characterization)

Let P be a distribution over N with finite multiinformation and u be a structural imset over N. Then the following conditions are equivalent:

(a) *P is Markovian with respect to u,*

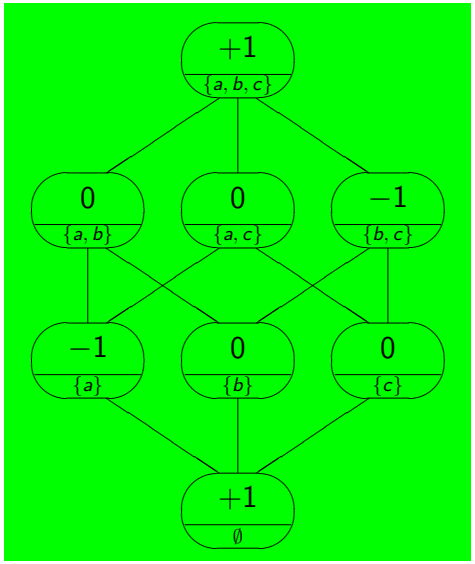
(b) *the marginal densities of P satisfy the product formula:*

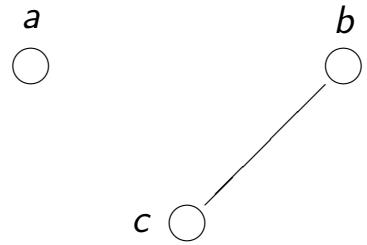
$$\prod_{A\subseteq N} f_A(x)^{u_+(A)} = \prod_{B\subseteq N} f_B(x)^{u_-(B)} \quad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N\,,$$

where u_+, u_- denote the positive and negative parts of u,

(c) $\langle m_P, u\rangle = 0$.

Example: illustration of the theorem





- $u := u_{\langle a,bc|\emptyset\rangle}$
- $u_+ = \delta_{abc} + \delta_\emptyset \quad u_- = \delta_a + \delta_{bc}$
- $f_\emptyset(x) \equiv 1 \quad m_P(\emptyset) = 0$

The following conditions are equivalent:

(a) $a \perp\!\!\!\perp bc \,|\, \emptyset \; [P]$,

(b) $f_{abc}(x) = f_a(x) \cdot f_{bc}(x)$ for μ-a.e. $x \in \mathsf{X}_N$,

(c) $m_P(abc) - m_P(a) - m_P(bc) = 0$.

Completeness of structural imsets

Theorem (completeness result)

Let P be a distribution over N with finite multiinformation.
Then there exists a structural imset u over N such that
P is perfectly Markovian with respect to u, that is, $\mathcal{M}_u = \mathcal{M}_P$.

In particular, every discrete CI structure and every regular Gaussian CI structure can faithfully be described by a structural imset.

Therefore, this linear-algebraic method overcomes the limitation of graphical approaches.

Structural imsets and supermodular functions

Definition (supermodular function)

A *supermodular function* is a real set function $m : \mathcal{P}(N) \to \mathbb{R}$ such that

$$m(A \cup B) + m(A \cap B) - m(A) - m(B) \geq 0 \quad \text{for all } A, B \subseteq N\,.$$

An equivalent definition is that $\langle m, v \rangle \geq 0$ for every $v \in \mathcal{E}(N)$.

Submodular functions, which play crucial role in optimization, are defined by converse inequalities ($=$ are (-1)-multiples of supermodular ones).

Theorem (supermodular characterization of structural imsets)

An imset u over N is structural iff it is o-standardized, which means that

- $\sum_{S\,:\,S \subseteq N} u(S) = 0$,
- *for any $i \in N$, one has* $\sum_{S\,:\,i \in S \subseteq N} u(S) = 0$,

and $\langle m, u \rangle \geq 0$ for any supermodular function m on $\mathcal{P}(N)$.

Independence implication

Definition (independence implication)

Suppose $u, v \in \mathcal{S}(N)$ are structural imsets. We say that *u independence implies v* and write $u \rightharpoonup v$ if $\mathcal{M}_v \subseteq \mathcal{M}_u$.

Lemma (independence implication characterization)

Provided that $u, v \in \mathcal{S}(N)$ one has $u \rightharpoonup v$ iff

$$\exists\, k \in \mathbb{N} \text{ such that } k \cdot u - v \in \mathcal{S}(N)\,.$$

Another equivalent condition is that, for *every supermodular* function m on $\mathcal{P}(N)$, one has $\langle m, v \rangle > 0 \;\Rightarrow\; \langle m, u \rangle > 0$.

(Standardized) supermodular functions form a (pointed) polyhedral cone in $\mathbb{R}^{\mathcal{P}(N)}$ and can be characterized finitely many generators of its *extreme* rays.

Therefore, to **disprove** $u \rightharpoonup v$ it suffices to find an *extreme supermodular function* m such that $\langle m, u \rangle = 0$ and $\langle m, v \rangle > 0$.

Testing CI implications by LP tools

The independence implication for imsets allow one to introduce the so-called *structural implication between CI statements*:

Given an input list L of CI statements over N and another CI statement t over N, we put $u_L := \sum_{\ell \in L} u_\ell$ (= the sum of semi-elementary imsets) and, if $u_L \rightharpoonup u_t$ then we say that L *structurally implies* t and write $L \Vdash t$.

This CI implication is stronger than the probabilistic one.

Moreover, the structural implication can effectively be tested by computational tools. One can expect that one needs to characterize extreme supermodular functions for this purpose, but it is not the case !

The point is that one can test structural implication using the methods of *linear programming* (LP).

R. Bouckaert, R. Hemmecke, S. Lindner, M. Studený (2010).
Efficient algorithms for conditional independence inference.
Journal of Machine Learning Research 11, 3453-3479.

Translation of Bayesian networks into imsets

The most popular graphical models in the area of probabilistic reasoning are *Bayesian networks* (BNs). The main idea of the algebraic approach here it to decribe the BN structure by a **unique vector representative**.

Definition (standard imset)

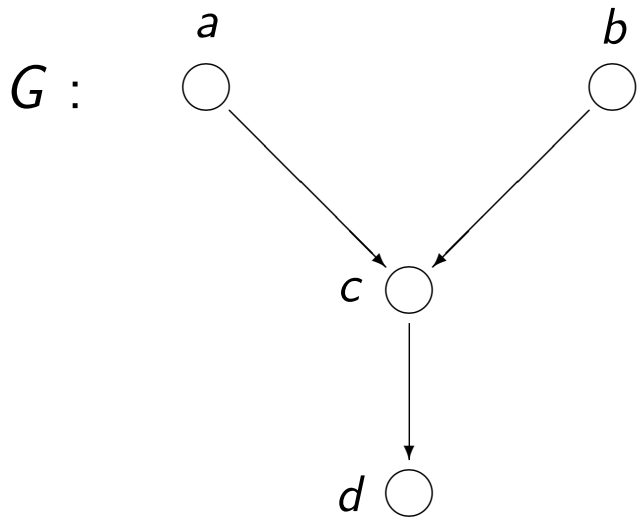
Given an acyclic directed graph G over N, the *standard imset* for G is given by the formula:

$$u_G := \delta_N - \delta_\emptyset + \sum_{i \in N} [\, -\delta_{\{i\} \cup pa_G(i)} + \delta_{pa_G(i)} \,],$$

where $pa_G(i) := \{ j \in N : j \to i \text{ in } G \}$ is the set of *parents* of i in G.

The case of a *decomposable graphical model* (= described by a chordal undirected graph H) can be viewed as a special case of a BN model.

Example: standard imset



$$
\begin{aligned}
u_G &= \delta_{\{a,b,c,d\}} - \delta_{\emptyset} + [\,-\delta_{\{a\}} + \delta_{\emptyset}\,] + [\,-\delta_{\{b\}} + \delta_{\emptyset}\,] \\
&\quad + [\,-\delta_{\{a,b,c\}} + \delta_{\{a,b\}}\,] + [\,-\delta_{\{c,d\}} + \delta_{\{c\}}\,] \\
&= \delta_{\emptyset} - \delta_{\{a\}} - \delta_{\{b\}} + \delta_{\{c\}} + \delta_{\{a,b\}} - \delta_{\{c,d\}} - \delta_{\{a,b,c\}} + \delta_{\{a,b,c,d\}} \\
&= u_{\langle a,b|\emptyset\rangle} + u_{\langle d,ab|c\rangle}
\end{aligned}
$$

Standard imsets and equivalence classes of BNs

Definition (independence equivalence of graphs)

We say that two graphs G and H over N are *independence equivalent* if $\mathcal{M}_G = \mathcal{M}_H$, where $\mathcal{M}_G := \{ \langle A, B|C\rangle : A \perp\!\!\!\perp B \,|\, C \ [G] \}$ denotes the CI model assigned to G by means of the respective separation criterion.

This typically means that the graphs are *Markov equivalent*, that is, they delimit the same class of Markovian distributions.

As concerns the Bayesian networks, we have this:

Lemma (uniqueness of the standard imset)

Acyclic directed graphs G and H over N are independence equivalent if and only if $u_G = u_H$.

Structural learning by maximizing a quality criterion

Most of learning methods are based on *maximizing a quality criterion*.

This is a real function $\mathcal{Q}$ of a graph G and observed database D which evaluates how good the graphical model given by G is to explain the occurrence of D.

The point is that every "reasonable" criterion $\mathcal{Q}$ for learning BN structures (e.g. BIC) becomes an affine function of the standard imset. Specifically:

$$\mathcal{Q}(G, D) = s_D^{\mathcal{Q}} - \langle t_D^{\mathcal{Q}}, u_G \rangle \,, \qquad \text{where } s_D^{\mathcal{Q}} \in \mathbb{R} \text{ and } t_D^{\mathcal{Q}} \in \mathbb{R}^{\mathcal{P}(N)}.$$

The vector $t_D^{\mathcal{Q}}$ is named the *data vector* (relative to $\mathcal{Q}$).

An important geometric observation is that the set of standard imsets over a fixed set of variables N is the *set of vertices (= extreme points) of a certain polytope* P. To apply efficient LP methods one needs to characterize the domain P in terms of linear inequalities in $\mathbb{R}^{\mathcal{P}(N)}$.

Conclusions

A lot of research effort has been devoted to attempts to characterize the considered polytopes in terms of linear inequalities.

For example, it was found that *extreme supermodular functions* correspond to important facet-defining inequalities for these polytopes.

J. Cussens, M. Studený, D. Haws (2017). Polyhedral aspects of score equivalence in Bayesian network structure learning. Mathematical Programming A 164(1/2), 285-324.

An elegant (and challenging) conjecture about the complete facet list of a polytope for learning decomposable models was recently disproved.

M. Studený, J. Cussens, V. Kratochvíl (2021). The dual polyhedron to the chordal graph polytope and the rebuttal of the chordal graph conjecture. International Journal of Approximate Reasoning 138, 188-203.