On non-graphical description of models of conditional independence structure

Milan Studený*

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4, 18208 Prague, Czech Republic
and also
Laboratory of Intelligent Systems
University of Economics Prague
E-mail: studeny@utia.cas.cz

Abstract

Several graphical structural models, including some models with latent variables can be viewed as models of conditional independence structure. However, usual graphical methods do not allow one to describe all possible stochastic conditional independence structures. Therefore an attempt to develop a general method of mathematical description of conditional independence structures by means of certain integer-valued functions, called *structural imsets*, was made. The main part of the paper is an outline of this approach. The presented results concern the mathematical basis of this method. After exposition of theoretical background some open questions are discussed: the problem of internal computer representation of models, inferential problems and interpretation question. The paper is concluded by a cursory reflection on what are suitable learning strategies and relevant data generating procedures for models of conditional independence structure.

1 Introduction

The mere aim of this introductory section is to explain motivation of research on *conditional independence* (CI) structures from the point of view of the purpose of the workshop. Some of structural models common in mathematical statistics can be interpreted as models of CI structure. Examples of such models are structural equation models (see Section 5 in [2]) and LISREL models (see Section 4 in [9]). In these models (non-degenerate) Gaussian random variables some of which are observable and some of which are latent (= hidden) are related by means of a system of linear equations. The point is that the class of possible joint distributions of observable variables can be often equivalently characterized as the class of (non-degenerate multidimensional) Gaussian measures satisfying certain CI statements. Thus, these models can be vieved as models of CI structure. A simple illustrative example is given in the next section (Example 2.1). Since detailed inspection of above mentioned statistical models is not the goal of this paper the reader is adviced to refer to [2, 9] for more details. I mentioned these models mainly in order to relate the topic of my contribution to central theme of the workshop.

Models of CI structure occur in other branches of mathematical statistics, in particular in discrete statistics dealing with categorical data. For example, well-known graphical log-linear models used in analysis of contingency tables (see Chapter 4 of [12]) can be interpreted as models of CI structure. Another area where stochastic CI is successfully applied is probabilistic reasoning [18] which is a branch of artificial intelligence in which decision-making under uncertainty is done on probabilistic basis. Note that the concept of CI is not exclusively a probabilistic concept. It was introduced in several non-probabilistic frameworks, namely in various calculi for dealing with uncertainty in artificial intelligence - for overview see [27, 20, 7].

*This work has been supported by the grant MŠMT n. VS96008 and by the grant GAČR n. 201/01/1482.

Traditional tools for description of CI structures are graphs whose nodes correspond to (random) variables. One can distinguish two traditional approaches: using undirected graphs named also 'Markov networks' and using acyclic directed graphs named also 'Bayesian networks' (this terminology originates from [18]). *Chain graphs* were introduced in mid-eighties [10] in order to provide an unifying point of view on these two traditional graphical approaches (for a more detailed exposition of chain graph models see [12]). Moreover, a lot of advanced graphical approaches to description of stochastic CI structures have occured in the past few years - for an overview see Chapter 3 of [33]. Note that some of these approaches use specific graphs for description of models involving latent variables.

Nevertheless, classic graphical approaches do not have potential for description of all possible stochastic CI structures induced by discrete probability measures. For example, the number of chain graph models over four variables is 200 while the number of all (discrete stochastic) CI structures over four variables is 18300 (see recently finished series of papers [13, 14, 15] or a survey paper [29]). The reason behind this phenomenon is that the number of discrete stochastic CI structures grows superexponentially with the number of variables while the number of graphs usually grows only exponentially with the number of nodes (= variables).

Of course, the reader may object that graphical models have a big advantage of easy interpretation and one can limit attention to a certain class of 'nice' graphical models. This attitude may later result in uncritical standardization of a certain class of graphical model, for example Bayesian networks in the area of probabilistic reasoning. Undesirable consequence such of development is that some people start to regard these models as all feasible and only acceptable models of CI structure and this may lead to serious methodological errors in learning procedures. For example, a lot of algorithms for learning Bayesian networks use 'edge removal' procedures. That means one uses statistical tests of CI statements to reveal CI structure and every positive result of such a test is represented by removal of a certain edge in a hypothetical graph (which should represent the searched CI structure). This procedure is completely correct in case that the CI structure induced by the underlying probability distribution (= distribution which 'generates' data) is precisely graphical CI structure. However, this seems to be quite rare event in light of the above mentioned comparison of number of models in case of four variables. The edge removal always represents acceptance of a new graphical model. From the point of view of learning CI structure it means that one accepts a whole collection of CI statements which are represented in the new graph (but were not represented in the old one). Thus, limitation to a certain graphical framework forces acceptance of additional CI statements (possibly not supported by data) on basis of a CI statement correctly recognized in data! The reason behind this is the discrepancy between accepted graphical framework and much wider framework of stochastic CI structures. Note that the above mentional error was already criticized in literature - see for example [34]. Needless to say that an analogous error can occur within other graphical approaches.

A safe way how to prevent repeated 'temptation' of elegant (but limited) classes of special models is to provide a method for description of *all* stochastic CI structures by means of objects of discrete mathematics. As explained above such a method has to use more complex mathematical tools than graphs. An attempt at such a method is the approach outlined in Section 3 which uses so-called *'structural imsets'*. The approach was first presented in the series of papers [30, 28] but the original exposition was complicated by superfluous technicalities and lack of transparent motivation. More up-to-date full version of mathematical basis of this approach (which includes formerly omitted continuous case as well) is in preparation [33]. The aim of this paper is both to outline the theoretical background of the method (of structural imsets) and to indicate research directions towards possible practical applicability of this approach.

Next section recalls basic concepts, in particular the definition of *conditional independence*. Section 3 is a survey of basic definitions and results of theory of structural imsets. In Section 4 (open) questions motivated by practical requirements are discussed. The problem of identification of 'data generating process' is briefly mentioned in Conclusion (Section 5).

2 Basic concepts

Throughout the paper the symbol N denotes a non-empty finite set of *variables*. Intended interpretation is that the variables correspond to primitive factors described by random variables. The power set of N will be denoted by $\mathcal{P}(N) \equiv \{A;\, A \subseteq N\}$. The following convention will be used in the sequel: given $A, B \subseteq N$ the juxtaposition AB denotes the union $A \cup B$. Moreover, the next symbols will be reserved for number sets: $\mathbf{R}$ denotes *real numbers*, $\mathbf{Z}$ *integers*, $\mathbf{Z}^+$ *non-negative integers* (including 0) and $\mathbf{N}$ *natural numbers* (without 0). The symbol $|A|$ will be used to denote the number of elements of a finite set A, that is its *cardinality*.

Definition 2.1 Basic notion is *probability measure over* N. This phrase describes the situation when a measurable space $(\mathsf{X}_i, \mathcal{X}_i)$ is given for every $i \in N$ and a probability measure P is defined on the Cartesian product $(\prod_{i\in N} \mathsf{X}_i, \prod_{i\in N} \mathcal{X}_i)$ of measurable spaces. Then $(\mathsf{X}_A, \mathcal{X}_A)$ will be used as a shorthand for $(\prod_{i\in A} \mathsf{X}_i, \prod_{i\in A} \mathcal{X}_i)$ for every $\emptyset \neq A \subseteq N$. Recall that the *marginal* of P for $\emptyset \neq A \subset N$, denoted by P^A is defined by the formula

$$P^A(\mathsf{A}) = P(\mathsf{A} \times \mathsf{X}_{N\setminus A}) \qquad \text{for } \ \mathsf{A} \in \mathcal{X}_A\,.$$

Moreover, put $P^N \equiv P$ and accept a formal convention that the marginal of P for $A = \emptyset$ is a probability measure on a (fixed appended) measurable space $(\mathsf{X}_\emptyset, \mathcal{X}_\emptyset)$ with trivial σ-algebra $\mathcal{X}_\emptyset = \{\emptyset, \mathsf{X}_\emptyset\}$. Observe that such a measurable space admits only one probability measure $P^\emptyset$ and that $(\mathsf{X}_\emptyset, \mathcal{X}_\emptyset) \times (\mathsf{X}_A, \mathcal{X}_A)$ is isomorphic to $(\mathsf{X}_A, \mathcal{X}_A)$ for every $\emptyset \neq A \subseteq N$. $\diamondsuit$

To give the definition of conditional independence within this framework one needs specific understanding of the concept of conditional probability. Given a probability measure P over N and disjoint sets $A, C \subseteq N$ by *conditional probability on* X_A *given* C will be understood a function of two arguments $P_{A|C} : \mathcal{X}_A \times \mathsf{X}_C \to [0,1]$ which ascribes to every $\mathsf{A} \in \mathcal{X}_A$ a $\mathcal{X}_C$-measurable function $P_{A|C}(\mathsf{A}|*)$ such that

$$P^{AC}(\mathsf{A} \times \mathsf{C}) = \int_{\mathsf{C}} P_{A|C}(\mathsf{A}|x)\; dP^C(x) \qquad \text{for every } \ \mathsf{C} \in \mathcal{X}_C\,.$$

Formally, a *conditional independence statement over* N is a statement of the form "A is conditionally independent of B given C" where $A, B, C \subseteq N$ are pairwise disjoint subsets of N. Such a statement should be always understood with respect to a certain mathematical object $\boldsymbol{o}$ over N, for example a probability measure over N. We will use the notation $A \perp\!\!\!\perp B \,|\, C\ [\boldsymbol{o}]$ then, but the symbol $[\boldsymbol{o}]$ can be omitted when it is suitable.

Thus, every conditional independence statement corresponds to a *disjoint triplet over* N, that is a triplet $\langle A, B|C\rangle$ of pairwise disjoint subsets of N. Here, punctation anticipates the intended role of component sets. The third component, put after the strait line, is the conditioning set, while two former components are independent areas, usually interchangeable. Formal difference is that such a triplet can be interpreted either as the corresponding independence statement, or (alternatively) as its negation, that is the corresponding *dependence statement*. The class of all disjoint triples over N will be denoted by $\mathcal{T}(N)$.

Definition 2.2 Given a probability measure P over N and a disjoint triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ one says that A is *conditionally independent of* B *given* C *with respect to* P if for every $\mathsf{A} \in \mathcal{X}_A$ and $\mathsf{B} \in \mathcal{X}_B$

$$P_{AB|C}(\mathsf{A} \times \mathsf{B}|x) = P_{A|C}(\mathsf{A}|x) \cdot P_{B|C}(\mathsf{B}|x) \qquad \text{for } \ P^C\text{-almost every } \ x \in \mathsf{X}_C\,. \tag{1}$$

Then writes $A \perp\!\!\!\perp B \,|\, C\ [P]$. $\diamondsuit$

Observe that in case $C = \emptyset$ it reduces to simple equality $P^{AB}(\mathsf{A} \times \mathsf{B}) = P^A(\mathsf{A}) \cdot P^B(\mathsf{B})$ which gives classic concept of stochastic independence. The validity of (1) does not depend on the choice of versions of conditional probabilities given C since these are determined uniquely just within equivalence P^C-almost everywhere. The definition above resembles general definition of concept of CI for σ-algebras - see for example [16]. However, an equivalent definition which is easy to verify is needed for practical purposes. One can introduce such an equivalent definition for a special class of probability measures.

Definition 2.3 A probability measure P on $(\mathsf{X}_N, \mathcal{X}_N)$ is *marginally continuous* if there exists a collection of σ-finite measures μ_i on $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ such that P is absolutely continuous with respect to the product measure $\mu \equiv \prod_{i\in N} \mu_i$, that is $P \ll \mu$. The measure μ will be called *dominating measure* for P. Having fixed μ and given $\emptyset \neq A \subseteq N$ every version of the Radon-Nikodym derivative of P^A with respect to $\mu_A \equiv \prod_{i\in A} \mu_i$ will be called the *marginal density* (of P) for A and denoted by f_A. By convention marginal density for the empty set is a constant function on $\mathsf{X}_\emptyset$ having value 1, that is $f_\emptyset \equiv 1$. ♢

The following lemma gives a simple equivalent definition of marginal continuity which perhaps explains my terminology. It can be verified using Proposition 1 in [24].

Lemma 2.1 A probability measure P over N is marginally continuous iff it is absolutely continuous with respect to the product of its one-dimensional marginals, that is $P \ll \prod_{i\in N} P^{\{i\}}$.

Marginal density for $A \subseteq N$ is a $\mathcal{X}_A$-measurable function on X_A. The following notation allows one to understand it as a function on X_N as well. Given $\emptyset \neq A \subseteq N$ and $x \in \mathsf{X}_N$ the symbol x_A will denote the *projection of x onto A*, that is $x_A = [x_i]_{i\in A}$ whenever $x = [x_i]_{i\in N}$. Moreover, the convention concerning the marginal density for $\emptyset$ means that in subsequent formulas one has $f_\emptyset(x_\emptyset) \equiv 1$ for every $x \in \mathsf{X}_N$.

Lemma 2.2 Let P be a marginally continuous measure over N. Then, for every $\langle A, B|C\rangle \in \mathcal{T}(N)$ one has $A \perp\!\!\!\perp B \,|\, C$ $[P]$ iff the following equality holds

$$f_{ABC}(x_{ABC}) \cdot f_C(x_C) = f_{AC}(x_{AC}) \cdot f_{BC}(x_{BC}) \qquad \text{for } \mu\text{-almost every } x \in \mathsf{X}_N. \tag{2}$$

The validity of (2) trivially does not depend on the choice of (versions) of marginal densities. The point of Lemma 2.2 is that it even does not depend on the choice of the dominating measure μ since $A \perp\!\!\!\perp B \,|\, C$ $[P]$ does depend on it as well!

Let me point out two typical special cases. In *discrete case* when X_i is a non-empty finite set every probability measure over N is marginally continuous since the counting measure can serve as a dominating measure. Every function $p : \mathsf{X}_N \to [0,1]$ such that $\sum\{p(x)\,;\, x \in \mathsf{X}_N\} = 1$ can serve as density of a probability measure on X_N then. The marginal density for $\emptyset \neq A \subset N$ is obtained by the formula

$$p_A(y) \;=\; \sum \{\, p(x,y)\,;\;\; x \in \mathsf{X}_{N\setminus A}\,\} \qquad \text{for every } \;\; y \in \mathsf{X}_A\,,$$

and (2) takes the form

$$p_{ABC}(x_{ABC}) \cdot p_C(x_C) = p_{AC}(x_{AC}) \cdot p_{BC}(x_{BC}) \qquad \text{for every } x \in \mathsf{X}_N. \tag{3}$$

In *Gaussian case* when X_i is $\mathbf{R}$ for each $i \in N$ every non-degenerate Gaussian measure over N is marginally continuous since the Lebesgue measure on $\mathbf{R}^N$ can serve as a dominating measure. Density f of the Gaussian measure $\mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})$ where $\boldsymbol{e} \in \mathbf{R}^N$ and $\boldsymbol{\Sigma}$ is a positive definite $N \times N$ real matrix is then

$$f(x) = \frac{1}{\sqrt{(2\pi)^{|N|} \cdot \det(\boldsymbol{\Sigma})}} \cdot \exp^{-\frac{(x-\boldsymbol{e})^\top \cdot \boldsymbol{\Sigma}^{-1} \cdot (x-\boldsymbol{e})}{2}} \qquad \text{for } x \in \mathbf{R}^N\,.$$

The marginal density for $\emptyset \neq A \subseteq N$ is given by the same formula where $\boldsymbol{e}$ is replaced by $\boldsymbol{e}_A$ and $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j\in N}$ by its main submatrix $\boldsymbol{\Sigma}_{A\cdot A} = (\sigma_{ij})_{i,j\in A}$. Speciality of Gaussian case is that CI statements $A \perp\!\!\!\perp B \,|\, C$ can be simply characterized in terms of conditional covariance matrix

$$\boldsymbol{\Sigma}_{AB|C} = \boldsymbol{\Sigma}_{AB\cdot AB} - \boldsymbol{\Sigma}_{AB\cdot C} \cdot (\boldsymbol{\Sigma}_{C\cdot C})^{-1} \cdot \boldsymbol{\Sigma}_{C\cdot AB}$$

which is also named Schur complement in matrix calculus. Indeed, one has $A \perp\!\!\!\perp B \,|\, C$ $[\mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})]$ iff (see Section 2.8 in [33]).

$$(\boldsymbol{\Sigma}_{AB|C})_{ij} = 0 \qquad \text{for every } i \in A \text{ and } j \in B\,. \tag{4}$$

Remark **2.1** Several authors independently drew attention to basic formal properties of conditional independence. In modern statistics, they were first accentuated by Dawid [6], then mentioned by Mouchart and Rolin [16].Spohn [23] interpreted them in the context of philosophical logic. Finally, their significance in probabilistic reasoning was discerned and highlighted by Pearl and Paz [17]. Their terminology [18] was later widely accepted, so that researchers in artificial intelligence started to call them the *semi-graphoid properties*.

1. $A \perp\!\!\!\perp \emptyset \,|\, D\ [P]$ — triviality,
2. $A \perp\!\!\!\perp B \,|\, D\ [P] \ \Rightarrow\ B \perp\!\!\!\perp A \,|\, D\ [P]$ — symmetry,
3. $A \perp\!\!\!\perp B \cup C \,|\, D\ [P] \ \Rightarrow\ A \perp\!\!\!\perp C \,|\, D\ [P]$ — decomposition,
4. $A \perp\!\!\!\perp B \cup C \,|\, D\ [P] \ \Rightarrow\ A \perp\!\!\!\perp B \,|\, C \cup D\ [P]$ — weak union,
5. $\{\, A \perp\!\!\!\perp B \,|\, C \cup D\ [P]\ \&\ A \perp\!\!\!\perp C \,|\, D\ [P] \,\} \ \Rightarrow\ A \perp\!\!\!\perp B \cup C \,|\, D\ [P]$ — contraction. □

Definition 2.4 By *conditional independence model* induced by a probability measure P over N will be understood the class of disjoint triplets over N

$$\mathcal{M}_P = \{\, \langle A, B|C\rangle \in \mathcal{T}(N);\ \ A \perp\!\!\!\perp B \,|\, C\ [P] \,\}\,.$$

By *formal independence model* over N will be understood any class $\mathcal{M} \subseteq \mathcal{T}(N)$. This phrase indicates that the involved triplets are interpreted as independence statements, although from purely mathematical point of view it is nothing but a subset of $\mathcal{T}(N)$. ♢

Remark **2.2** This is to explain some terminological peculiarities. The word 'model' can be understood in a few different ways. This happens even in this paper. The first meaning (used mainly in the introductory section) is a general denomination for arbitrary consistent collection of ideas and assumptions describining certain situation. Adjective 'structural' then means that the described situation exhibits certain structural relationships. Particular examples of such models are miscellaneous models used in mathematical statistics.

Majority of these models can be formulated as an assumption that an unknown probability distribution which 'generates' data belongs to a certain class of (multidimensional) probability distributions. Thus, from mathematical point of view *statistical model* can be identified with a class of distributions. This is the second possible undestanding of the word 'model' which can be found in some textbooks of mathematical statistics which put emphasis on an abstract point of view. The phrase 'graphical model' is then used if the class of distributions is somehow determined by a graph whose nodes correspond to variables.

The third meaning of the word 'model' is a mathematical notion *independence model* introduced in the definition above. Note that I loosely follow terminology given by Pearl [18] here. Well, every independence model in this sense can be identified with a statistical model as follows. Given a certain basic class of distributions $\mathcal{L}$ (say the class of discrete measures with fixed sample space X_N or the class of non-degenerate Gaussian measures over N) every collection of disjoint triplets over N determines a subclass of $\mathcal{L}$ composed of distributions satisfying respective CI statements. In other words, it can be identified with a statistical model which is called then the 'model of CI structure'. □

Example 2.1 This example illustrates that some common statistical models can be viewed as models of CI structure. Consider the following (recursive) linear structural equation model with uncorrelated errors (I follow terminology from [2], Section 5):

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= \varepsilon_2 \\ X_3 &= a \cdot X_1 + b \cdot X_2 + \varepsilon_3 \\ X_4 &= c \cdot X_2 + \varepsilon_4 \\ X_5 &= d \cdot X_3 + e \cdot X_4 + \varepsilon_5 \end{aligned} \tag{5}$$

where a, b, c, d, e are real parameters and error variables $\varepsilon_1, \ldots, \varepsilon_5$ are supposed to be independent Gaussian variables with zero mean (their variances are additional free parameters). Then the vector of observed variables $\boldsymbol{X} = [X_1, \ldots, X_5]^\top$ can be expressed as a linear function $\boldsymbol{X} = \boldsymbol{A} \cdot \boldsymbol{E}$

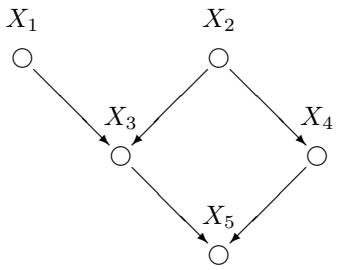

Figure 1: Path diagram corresponding to the structural equation model (5).

of a random vector $\boldsymbol{E} = [\varepsilon_1, \ldots, \varepsilon_5]^\top$ where $\boldsymbol{A}$ is a lower triangular matrix depending on the parameters. Since $\boldsymbol{A}$ is regular one can deduce that $\boldsymbol{X}$ has non-degenerate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{A} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{A}^\top)$ where $\boldsymbol{\Gamma}$ is a diagonal matrix whose i-th diagonal element is the variance of ε_i $(i = 1, \ldots, 5)$. Suppose that one is interested in the question what are all possible distributions of random vectors $\boldsymbol{X}$ satisfying (5) for any choice of 10 free parameters (with proviso that variances of ε_i are non-zero). A method of *path diagram* (see [2], Section 5.1) allows one to ascribe a certain acyclic directed graph to the system (5), namely the graph in Figure 1. Further results reported in [2] then allow one to identify on basis of the diagram CI statements valid in every such distribution P. Basic observation is that the class of distributions of $\boldsymbol{X}$ coincides with the class of non-degenerate Gaussian measures P satisfying

$$\{1\} \perp\!\!\!\perp \{2\} \,|\, \emptyset \; [P] \qquad \{1,3\} \perp\!\!\!\perp \{4\} \,|\, \{2\} \; [P] \qquad \{1,2\} \perp\!\!\!\perp \{5\} \,|\, \{3,4\} \; [P] \,. \tag{6}$$

Well, I do not have a complete exact proof of this claim for a general path diagram (with possibly corelated errors). My belief is based on some calculation and the next indirect argument. The dimension of both classes of Gaussian measures (= the number of free paramenters) is the same and the class of measures satisfying (6) is known to be a subset of the other class. Thus, the model (5) can ve viewed as a statistical model of CI structure. □

3 Non-graphical approach

By an *imset over* N is understood an integer-valued function on the power set of N, that is any function $u : \mathcal{P}(N) \to \mathbf{Z}$, or alternatively an element of $\mathbf{Z}^{\mathcal{P}(N)}$. Basic operation with imsets, namely summation, subtraction, multiplication by an integer are defined coordinatewisely. *Multiset* is an imset with non-negative values, that is any function $m : \mathcal{P}(N) \to \mathbf{Z}^+$. Every imset u over N can be written as a difference $u = u^+ - u^-$ of two multisets over N where u^+ is the *positive part* of u and u^- is the *negative part* of u, defined as follows:

$$u^+(S) = \max\{u(S), 0\}\,, \quad u^-(S) = \max\{-u(S), 0\} \qquad \text{for} \;\; S \subseteq N\,.$$

The word 'multiset' is taken from combinatorial theory [1] while the word 'imset' is an abbreviation for **i**nteger-valued **m**ulti**set**.

It is clear how to represent an imset over N in memory of a computer, namely by a vector with $2^{|N|}$ integer components which correspond to subsets of N. For a small number of variables, one can also visualize imsets in a more telling way, using special pictures. The power set $\mathcal{P}(N)$ is a distributive lattice and can be represented in the form of *Hasse diagram* (see p. 6 in [3]). Nodes of this diagram correspond to elements of $\mathcal{P}(N)$, that is to subsets of N, and a link is made between two nodes if the symmetric difference of the represented sets is a singleton. A function on $\mathcal{P}(N)$ can be visualized then in such a way that one writes assigned values into respective nodes. For example, the imset u over $N = \{a, b, c\}$ defined by the table

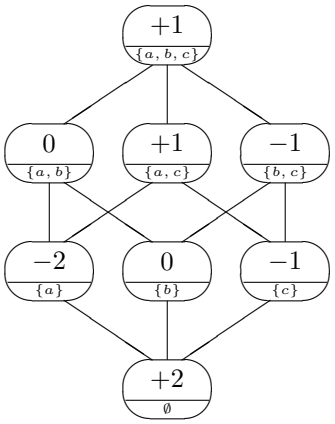

Figure 2: Hasse diagram of an imset over $N = \{a, b, c\}$.

S	$\emptyset$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a, b\}$	$\{a, c\}$	$\{b, c\}$	$\{a, b, c\}$
$u(S)$	$+2$	-2	0	-1	0	$+1$	-1	$+1$

can be visualized in the form of the diagram from Figure 2.

3.1 Structural imsets

However only a certain type of imsets is suitable for description of stochastic CI models.

Definition 3.1 Every disjoint triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ corresponds to a *semi-elementary imset* $u_{\langle A,B|C\rangle}$ defined as follows:

S	ABC	AC	BC	C	any other $S \subseteq N$
$u_{\langle A,B|C\rangle}(S)$	$+1$	-1	-1	$+1$	0

It is called *elementary* if A and B are singletons. An imset u over N is called *structural* if there exists a natural number $l \in \mathbf{N}$, a sequence of semi-elementary imsets $u_1, \ldots, u_r$, $r \geq 0$ and non-negative integers $k_i \in \mathbf{Z}^+$ such that

$$l \cdot u = \sum_{i=1}^{r} k_i \cdot u_i \,.$$

Given a structural imset u the *lower class* of u is the collection of sets

$$\mathcal{L}_u \,=\, \{\, S \subseteq N \,;\, \text{there exists } S \subseteq T \subseteq N \text{ with } u(T) < 0 \,\}\,,$$

and the *upper class* of u is the collection of sets

$$\mathcal{U}_u \,=\, \{\, S \subseteq N \,;\, \text{there exists } S \subseteq T \subseteq N \text{ with } u(T) > 0 \,\}\,,$$

The set $R_u = \bigcup\{S;\, S \in \mathcal{U}_u\}$ is called the *range* of u. Note that one always has $\mathcal{L}_u \subseteq \mathcal{U}_u$ and $R_u = \bigcup\{S;\, S \in \mathcal{L}_u\}$ (see Section 4.2.3 in [33]). Typically, $\mathcal{U}_u$ is strictly bigger than $\mathcal{L}_u$ (they coincide only in case $u = 0$). ♢

An example of a structural imset $u_{\langle a,b|c\rangle} + 2 \cdot u_{\langle a,c|\emptyset\rangle}$ is given in Figure 2. Thus, a structural imset can be viewed as an imset obtained as a conical combination of semi-elementary imsets (necessarily with rational coefficients). However, it can be equivalently introduced as an imset obtained as a conical combination of elementary imsets. There are theoretical reasons for this apparently superfluous terminological distinction: elementary imsets correspond (in sense of the following definition) to atomic (= minimal non-trivial) stochastic CI models.

Definition 3.2 Let u be a structural imset over N and $\langle A, B|C\rangle \in \mathcal{T}(N)$ a disjoint triplet over N. One says that $\langle A, B|C\rangle$ is represented in u and writes $A \perp\!\!\!\perp B \,|\, C \;[u]$ if there exists a natural number $k \in \mathbf{N}$ and a structural imset w such that

$$k \cdot u \;=\; u_{\langle A,B|C\rangle} + w\,.$$

The *CI model* induced u is then

$$\mathcal{M}_u = \{\, \langle A, B|C\rangle \in \mathcal{T}(N);\;\; A \perp\!\!\!\perp B \,|\, C \;[u]\,\}\,.$$

A probability measure P over N is *Markovian* with respect to a structural imset u if

$$A \perp\!\!\!\perp B \,|\, C \;[u] \;\;\text{implies}\;\; A \perp\!\!\!\perp B \,|\, C \;[P] \quad \text{for every } \langle A, B|C\rangle \in \mathcal{T}(N)\,.$$

It is called *perfectly Markovian* if the converse implication is true as well. ♢

Markov condition above is completely analogous to usual Markov conditions used in various graphical models [12]. The difference is that instead of a graphical separation criterion which is used to determine whether a disjoint triplet over N is represented in a given graph an algebraic criterion is introduced. The fact that P is perfectly Markovian with respect to u means that CI model induced by P is precisely described by u. In other words, u serves as an object of discrete mathematics which describes the CI structure hidden in P.

The following *coincidence principle* holds (see Consequence 4.3 in [33]). Markovian measures with respect to a structural imset u which have identical marginals for sets in $\mathcal{L}_u$ have identical marginals for sets in $\mathcal{U}_u$. This fact hopefully justifies the above accepted terminology 'lower' and 'upper' class. Observe that this principle works only within the range R_u of the imset.

3.2 Multiinformation

The class of probability measures for which this approach is applicable, that is whose induced CI models can be described by structural imsets, is relatively wide. It is the class of *measures with finite multiinformation* [24].

Definition 3.3 Suppose that $(\mathsf{X}, \mathcal{X})$ is a measurable space, P is a probability measure and μ is a σ-finite measure on $(\mathsf{X}, \mathcal{X})$ such that $P \ll \mu$. By *relative entropy of P with respect to μ* will be understood the integral

$$H(P\,|\,\mu) = \int_{\mathsf{X}} \ln \frac{dP}{d\mu}(x)\;\, dP(x) \equiv \int_{\mathsf{X}} \frac{dP}{d\mu}(x) \cdot \ln \frac{dP}{d\mu}(x)\;\, d\mu(x)\,,$$

provided that the function $\ln \frac{dP}{d\mu}$ is P-quasi-integrable. The *multiinformation* of a probability measure P over N is the relative entropy of P with respect to the product of its one-dimensional marginals $\prod_{i\in N} P^{\{i\}}$. In the case that $H(P|\prod_{i\in N} P^{\{i\}})$ is finite one can introduce the *multiinformation function* induced by P as follows:

$$m_P(A) \;=\; H(P^A \,|\, \prod_{i\in A} P^{\{i\}}) \qquad \text{for } A \subseteq N\,, \tag{7}$$

and $m_P(\emptyset) = 0$ by convention. ♢

Note that multiinformation is always non-negative but it can take the value $+\infty$ in general. In particular, every multiinformation function is non-negative. The class of measures with finite multiinformation is a subclass of the class of marginally continuous measures (see Section 2.3.4 of [33]). It is a rather wide class of measures since it includes typical measures used in practice.

Discrete measures These simple probability measures are mainly used in probabilistic reasoning [18]. Strictly positive discrete probability measures are also behind models used in analysis of contingency tables (see [12], Chapter 4). The fact that every discrete probability measure over N has finite multiinformation is trivial.

Non-degenerate Gaussian measures These measures are widely used in multivariate statistics [5]. It is well-known fact that every non-degenerate Gaussian measure over N has finite multiinformation (see e.g. [33]).

Non-degenerate conditional Gaussian measures This class of measures was proposed by Lauritzen and Wermuth [11] with the aim to unify frameworks for discrete and continuous graphical models. Non-degenerate conditional Gaussian measure P over N, called also shortly *CG-measure over* N also have finite multiinformation (see Consequence 4.1 in [33]).

3.3 Results

The main result is as follows (see Chapter 5 in [33], a discrete version [30]).

Theorem 3.1 *Let P be a probability measure over N with finite multiinformation. Then there exists a structural imset u over N such that P is perfectly Markovian with respect to u.*

Thus, every discrete stochastic CI structure can be described by a structural imset. On the other hand, there exists a structural imset inducing a formal independence model which is not stochastic CI model for any discrete probability measure.

Some readers may object that structural imsets are far from reasonable interpretation. Perhaps they appreciate the following equivalent definition of Markovian distribution in terms of certain product formula (see Chapter 4 of [33], discrete version [28]).

Theorem 3.2 *Let u be a structural imset over N, P a probability measure on $(\mathsf{X}_N, \mathcal{X}_N)$ with finite multiinformation. Suppose that $\mu = \prod_{i\in N} \mu_i$ is a dominating measure for P such that $-\infty < H(P^{\{i\}} \,|\, \mu_i) < \infty$ for every $i \in N$. Then the following two conditions are equivalent to the requirement that P is Markovian with respect to u.*

(i) $\prod\limits_{S\subseteq N} f_S(x_S)^{u^+(S)} = \prod\limits_{S\subseteq N} f_S(x_S)^{u^-(S)}$ *for μ-almost every $x \in \mathsf{X}_N$,*

(ii) $\sum_{S\subseteq N} m_P(S)\cdot u(S) = 0$ *where m_P is the multiinformation function from Definition 3.3.*

Note that in discrete case the formula (i) reduces to the form

$$\prod_{S\subseteq N} p_S(x_S)^{u^+(S)} = \prod_{S\subseteq N} p_S(x_S)^{u^-(S)} \quad \text{for every } x \in X_N\,.$$

The formula (i) also illustrates 'coincidence principle' mentioned above. Its left-hand side depends on the marginals for sets in the upper class $\mathcal{U}_u$ while its right-hand side depends on the marginals for sets in the lower class $\mathcal{L}_u$. The condition (ii) enables one to look at Markovian measures from the point of view of information theory (through the concept of multiinformation function). This elegant identity also brings taste of algebra in this area.

In order to determine the CI model induced by a structural imset (see Definition 3.2) one should be able to decide whether an imset is structural or not. Further result gives a certain theoretical solution to this task. These are auxiliary concepts.

Definition 3.4 An imset u over N is *normalized* if the collection of numbers $\{u(S);\, S \subseteq N\}$ has no common prime divisor. It is called *o-standardized* if

$$\sum_{S\subseteq N} u(S) = 0 \quad \text{and} \quad \forall\, i \in N \sum_{S\subseteq N,\, i\in S} u(S) = 0\,.$$

A function $m : \mathcal{P}(N) \to \mathbf{R}$ is *l-standardized* if

$$m(S) = 0 \qquad \text{whenever} \;\; S \subseteq N,\; |S| \leq 1\,,$$

For every pair m, u of real functions on $\mathcal{P}(N)$ introduce their *scalar product* by

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

A function $m : \mathcal{P}(N) \to \mathbf{R}$ is *supermodular* if $\langle m, u\rangle \geq 0$ for every semi-elementary imset u.

A supermodular function m can be equivalently introduced by the requirement $\langle m, u\rangle \geq 0$ for every elementary imset u. Note that every l-standardized supermodular imset is necessarily a non-negative function, therefore a multiset.

Theorem 3.3 *Given a non-empty finite set of variables N there exists the least finite collection of l-standardized normalized multisets $\mathcal{S}$ such that structural imsets over N are characterized as follows. An imset u over N is structural iff it is o-standardized and*

$$\langle m, u\rangle \geq 0 \qquad \textit{for every} \;\; m \in \mathcal{S}\,.$$

The proof of this result can be found in slightly modified (cryptic) form in [25] as Proposition 3. The main idea of the proof is simple. First, one has to verify that an imset u is structural iff it is o-standardized and $\langle m, u\rangle \geq 0$ for every supermodular function m. Then the cone of l-standardized supermodular functions is shown to have finitely many extreme rays. Each of these rays includes just one normalized multiset. The elements of $\mathcal{S}$ then correspond to the extreme rays of that cone.

Of course, the class $\mathcal{S}$ is determined by the requirement in Theorem 3.3 uniquely. I started to call $\mathcal{S}$ in [30] the *skeleton* (since it plays the role of the 'outer skeleton' of the cone of l-standardized supermodular functions: the cone can be obtained as its conical hull). The skeleton was found for $|N| \leq 5$ [32]. It has 5 elements if $|N| = 3$, 37 elements if $|N| = 4$ and 117978 elements if $|N| = 5$ (in which case it was found by means of a computer).

The concept of skeleton provides an elegant theoretical solution to the problem of characterization of equivalent structural imsets. Recall that (in the framework of graphical models) two graphs over N are supposed *Markov equivalent* (relative to a given fixed sample space $(\mathsf{X}_N, \mathcal{X}_N)$) if their classes of Markovian measures on $(\mathsf{X}_N, \mathcal{X}_N)$ coincide. Some authors prefer even stronger requirement that the graphs induce the same CI model (through respective graphical separation criterion) [12]. These two concepts are in fact equivalent (for non-trivial sample spaces). In fact, the difference can occur only if there is no perfect Markovian measure for one of two graphs. However, such a measure exists for every chain graph [31]. Nevertheless, as mentioned earlier structural imsets without perfectly Markovian measures exist which means that one has to distinguish two types of equivalence of structural imsets. Let me concentrate on the stronger equivalence.

Definition 3.5 Two structural imsets u, v over N are called *facially equivalent* if they induce the same CI model, that is $\mathcal{M}_u = \mathcal{M}_v$ (see Definition 3.2).

One says that a finite set of structural imsets L over N *facially implies* a structural imset u over N and writes $L \mapsto v$ if

$$\bigcup_{u \in L} \mathcal{M}_u \subseteq \mathcal{M}_w \qquad \text{implies} \;\; \mathcal{M}_v \subseteq \mathcal{M}_w$$

for every structral imset w over N. ♢

Evidently, facial equivalence of imsets u and v can be defined as mutual facial implication $u \mapsto v$ and $v \mapsto u$. The term 'facial' was introduced in [30]. It was inspired by the fact that the lattice of factor-classes of structural imsets (according to this equivalence) is isomorphic to a certain lattice of faces of a special polyhedron. The point is that facial implication can be characterized in terms of algebraic operations with integers (see Lemma 2.2 in [30], the 2nd part):

Lemma 3.1 Given a finite set L of structural imsets over N and a structural imset v over N one has $L \mapsto v$ iff there exist $k_w \in \mathbf{Z}^+$, $w \in L$ such that $\sum_{w \in L} k_w \cdot w - v$ is a structural imset.

Researchers in the area of graphical models were interested in graphical characterization of equivalent graphs [8]. The next consequence of Theorem 3.3 provides an elegant characterization of facial equivalence. It makes no problem to verify on basis of the preceding lemma and Theorem 3.3 the following result.

Consequence 3.1 Let $\mathcal{S}$ be the skeleton for N, L a finite set L of structural imsets and v a structural imset over N. Then $L \mapsto v$ iff

$$\forall\, m \in \mathcal{S} \quad \langle m, v\rangle > 0 \quad \Rightarrow \quad [\,\text{there exists } u \in L \text{ with } \langle m, u\rangle > 0\,]\,.$$

In particular, two structural imsets u, v are facially equivalent iff

$$\forall\, m \in \mathcal{S} \quad \langle m, v\rangle > 0 \quad \Leftrightarrow \quad \langle m, u\rangle > 0\,.$$

4 Discussion - research directions

Questions motivated by 'practice' are dealt with in this section.

4.1 Interpretation

Humans usually desire 'lucid' interpretation of accepted models. For example, in graphical models, single directed edges are sometimes interpreted as (perturbated) causal relationships among variables [22]. Therefore, the question what is interpretation of general model of CI structure is very natural.

I don't think that they have straightforward interpretation, at least not all of them. However, in my opinion, structural imsets indicate (through the product formula from Theorem 3.2) how reasonable interpretation of models of CI structures could be developed. The product formula looks to be a loose analogue of factorization defining log-linear models: the maximal sets of the lower class $\mathcal{L}_u$ play the role of a generating class of a log-linear model. The analogy is transparent in the case of well-known decomposable models [12]. Every such a model can be interpreted as a model of CI structure induced by the following structural imset (see Section 2.8 in [26])

$$u = \delta_{\cup\mathcal{C}} + \sum_{\emptyset\neq\mathcal{B}\subseteq\mathcal{C}} (-1)^{|\mathcal{B}|} \cdot \delta_{\cap\mathcal{B}} \tag{8}$$

where $\mathcal{C}$ is the class of cliques of respective chordal graph over N and δ_S is the 'indentificator' imset defined as follows

$$\delta_S(T) = \begin{cases} 1 & \text{if } T = S \\ 0 & \text{if } T \neq S \end{cases} \quad \text{for } T \subseteq N\,.$$

The respective product formula takes then the well-known form (see Section 2.8 [26])

$$p(x) \cdot \prod_{S\in\mathcal{D}} p_S(x_S)^{w(S)} = \prod_{S\in\mathcal{C}} p_S(x_S)$$

where $\mathcal{D}$ is the class of *separators* (see p. 15 in [12]) and $w(S)$ denotes the multiplicity of a separator S. The meaning of this formula is that the marginals on the cliques (= the maximal sets of the lower class) determine whole distribution (= the marginals on the upper class).

An open question is which models of CI structure admit such a clear interpretation.

4.2 Visualization

Related question is whether structural imsets can be represented by pictures. Hasse diagrams are suitable only for a small number of variables (at most 5) since they become complicated for higher number of variables.

I don't think that every structural imset over 6 and more variables can be simply visualized. Nevertheless, perhaps some 'nice' models (like decomposable models) whose respective structural imsets have 'a lot of zeros' can be visualized as Hasse diagrams of respective sub-posets (consisting of nodes with non-zero values). This sub-poset and the resulting Hasse diagram can be sometimes 'prunned' to represent essential information only. For example, in case of decomposable models (and the respective structural imset given by (8)) well-known *junction-trees* [4] or their modifications can be obtained in this way. Similar, but more general sparse hypergraphs called 'valuation networks' were used by Shenoy [20].

An open question is which models of CI structure admit such type of visualization.

4.3 Computer implementation

Theoretical way of implementation of 'facial' inference mechanism for structural imsets is clear (see Consequence 3.1). Practical problem is increasing cardinality of the skeleton with the number of variables. The grow seems to be superexponential. Several ways of overcoming the obstactle deserve attention.

The first chance consists in finding suitable theoretical characterization of all skeletal multisets (= elements of $\mathcal{S}$). There is a certain theoretical characterization of extremal supermodular set

functions [19] but this characterization does not seem to be suitable for the purpose of practical implementation. The desired hypothetical characterization may lead to an elegant classification and encoding of skeletal imsets (such a classification is possible in case $|N| = 4$). Then the skeletal multisets would not have been stored in memory of a computer but would have been recovered on basis of that classification when necessary.

The second chance is that testing facial implication for some special structural imsets (I have in mind those imsets for which the difference between the lower and upper class is not so big) does not seem to be very complex. It is quite probable that only some of skeletal multisets are essential for verification what is implied by such a structural imset.

The third chance is to base testing whether an imset is structural directly on Definition 3.1 and testing facial implication on Lemma 3.1. These together say that $u \mapsto v$ where u is structural and v elementary imset iff there exists $k \in \mathbf{N}$ such that $k \cdot u - v$ can be written in the form $\sum_{i=1}^{r} u_i$ where u_i are (possibly repeated) elementary imsets. In this case r can be determined on basis of $k \cdot u - v$ and elementary imsets are known as well. The question is what is the upper limit for the constant k. It certainly depends on N. This limit is 1 if $|N| \leq 4$ [25] and at least 7 if $|N| = 5$ [32].

4.4 Learning

The question of learning general CI structures was not studied very much. One can distinguish two types of learning:

- learning models of CI structure,
- learning particular probability distribution within given (= previously determined) model of CI structure.

Let me mention learning CI structures first. As explained in Section 1 structural imsets prevent some methodological errors which may occur if one is limited to a graphical framework. On the other hand, the state space of all CI structures is too big which leads to computational problems.

Every model of CI structure can be represented in memory of a computer by a structural imset. This idea leads directly to the question of choice of suitable representative within every class of facially equivalent structural imsets. This is an open problem. On the other hand, suitable candidates for representatives of some graphical models are known. For example, given an acyclic directed graph G the respective structural imset is given by

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

where $\pi(i)$ denotes the set of parents of a node $i \in N$. Note that two Markov equivalent graphs define the same imset by this formula.

Thus, particular methods of learning CI structures are to be developed. The theory of structural imsets can offer some heuristics. One of possible approaches is motivated by the condition (ii) in Theorem 3.2. One can try to estimate the multiinformation function m_P on basis of data and then to determine the respective structural imset u in such a way that $\langle m_P, u \rangle$ is close to zero (it is always non-negative).

The second topic of interest is learning particular probability distribution on condition that that the model of CI structure is already determined. This leads to the question what is 'natural' parametrization of the given model of CI structure. Of course, one is interested in such a way of parametrization in which parameters are 'independent each other'. Again, this is an open question which deserves detailed research.

Unlike the case of learning CI structures several particular methods for learning distribution within a given graphical model were developed (see e. g. [21]). Perhaps structural imsets can offer innovate heuristics also in this area. For example, the formula (i) in Theorem 3.2 may lead to an iterative method of estimation of underlying probability within certain models, namely those models in which the upper class is whole $\mathcal{P}(N)$. Then (i) can turn (by dividing) into a formula for joint density in terms of proper marginal densities. Such a formula then can form a basis of an iterative method (repeated substitution to the right-hand side of the formula gives the next iteration). An open question is what assumptions are needed to ensure convergence of such a method.

5 Conclusion

Let me conclude with a methodological consideration. The description of themes of the workshop includes an allusion to 'data generating process'. To be honest I do not understand completely what is general meaning of this notion (I am not a statistician by education). I can only estimate on basis of some parts of [5] that it is a collection of assumptions made by humans (= statisticians), namely that there are certain structural relationships among investigated factors (= variables) which are manifested by specific dependencies in data. The final goal of statistical analysis and interpretation is to reveal these 'causal relationships' which are usually described by means of a colection of (perturbated) functional relationships among variables.

Well, these functional relationships may manifest in data in the form of CI relationships which can possibly be indicated on basis of data. The revealed CI structure can be then used to 'find' the assumed functional relationships. However, this step in not based on empirical experience but on the assumptions made by humans! Therefore it is necessary to distingush between *causal interpretation* of graphical models (treated for example in [22]) and their *interpretation in terms od CI*. I think that one has to be aware of their difference and separate them strictly.

In my opinion, CI structure is something 'more empirical' and can be possibly learned 'automatically' on basis of data while causal explanation is human interpretation of empirical findings. Causal interpretation therefore cannot be performed 'automatically' and should be made by humans (practical statisticians or experts in the respective field). Thus, in my view, the question of finding suitable 'data generating process' should be formulated as a two-step task.

1. Find the most fitting CI structure to data (this represents *empirical finding*).
2. Ask which one of available models of 'data generating process' suits best the chosen CI structure (this represents *interpretation step*).

The criterion of choice of suitable data generating process is then given clearly by humans (with their full responsibility). Well, the second task then becomes an open mathematical problem which can be thoroughly studied.

Acknowledgements

I would like to express my thanks to my discussant Prof. J.-M. Rolin for his comments. I will try utilize them in a future publication which will be based on this overview paper.

References

[1] M. Aigner: Combinatorial Theory, Springer-Verlag, 1979.

[2] S. A. Andersson, D. Madigan, M. D. Perlman, T. S. Richardson: Graphical Markov models in multivariate analysis, in Multivariate Analysis, Design of Experiments, and Survey Sampling, Statistical Textbooks Monographs 159, Dekker 1999, pp. 189-229.

[3] G. Birkhoff: Lattice Theory, AMS Colloquim Publications, 1951.

[4] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter: Probabilistic Networks and Expert Systems, Springer-Verlag 1999.

[5] D. R. Cox, N. Wermuth: Multivariate Dependencies - Models, Analysis and Interpretation, Chapman and Hall 1996.

[6] A. P. Dawid: Conditional independence in statistical theory, Journal of the Royal Statistical Society series B 41 (1979), n. 1, pp. 1-31.

[7] A. P. Dawid: Separoids; a general framework for conditional independence and irrelevance, submitted to Annals of Mathematics and Artificial Intelligence.

[8] M. Frydenberg: The chain graph Markov property, Scandinavian Journal of Statistics 17 (1990), n. 4, pp. 333-353.

[9] J. T. A. Koster: Markov properties of nonrecursive causal models, Annals of Statistics 24 (1996), n. 5, pp. 2148-2177.

[10] S. L. Lauritzen, N. Wermuth: Mixed interaction models, research report R-84-8, Inst. Elec. Sys., University of Aalborg 1984.

[11] S. L. Lauritzen, N. Wermuth: Graphical models for associations between variables, some of which are qualitative and some quantitative, Annals of Statistics 17 (1989), n. 1, pp. 31-57.

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

[13] F. Matúš, M. Studený: Conditional independences among four random variables I., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 269-278.

[14] F. Matúš: Conditional independences among four random variables II., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 407-417.

[15] F. Matúš: Conditional independences among four random variables III., final conclusion, Combinatorics, Probability and Computing 8 (1999), n. 3, pp. 269-276.

[16] M. Mouchart, J.-M. Rolin: A note on conditional independence with statistical applications, Statistica 44 (1984), n. 4, pp, 557-584.

[17] J. Pearl, A. Paz: Graphoids, graph-based logic for reasoning about relevance relations, in Advances in Artificial Intelligence II (B. Du Boulay, D. Hogg, L. Steels eds.), North-Holland 1987, pp. 357-363.

[18] J. Pearl: Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference, Morgan Kaufmann 1988.

[19] J. Rosenmüller, H. G. Weidner: Extreme convex set functions with finite carrier: general theory, Discrete Mathematics 10, n. 3/4, pp. 343-382.

[20] P. P. Shenoy: Conditional independence in valuation-based systems, International Journal of Approximate Reasoning 10 (1994), n. 3, pp. 203-234.

[21] D. J. Spiegelhalter, S. L. Lauritzen: Sequential updating of conditional probabilities on directed graphical structures, Networks 20 (1990), n. 5, pp. 579-605.

[22] P. Spirtes, C. Glymour, R. Scheines: Causation, Prediction, and Search, Lecture Notes in Statistics 81, Springer-Verlag 1993.

[23] W. Spohn: Stochastic independence, causal independence and shieldability, Journal of Philosophical Logic 9 (1980), n. 1, pp. 73-99.

[24] M. Studený: Multiinformation and the problem of characterization of conditional independence relations, Problems of Control and Information Theory 18 (1989), n. 1, pp. 3-16.

[25] M. Studený: Convex set functions I. and II., research reports n. 1733 and n. 1734, Institute of Information Theory and Automation, Prague, November 1991.

[26] M. Studený: Multiinformation and conditional independence II., research report n. 1751, Institute of Information Theory and Automation, Prague, September 1992.

[27] M. Studený: Formal properties of conditional independence in different calculi of AI, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (M. Clarke, R. Kruse, S. Moral eds.), Lecture Notes in Computer Science 747, Springer-Verlag 1993, pp. 341-348.

[28] M. Studený, Description of conditional independence structures by means of imsets: a connection with product formula validity, in Uncertainty in Intelligent Systems (B. Bouchon-Meunier, L. L. Valverde and R. R. Yager eds.), Elsevier 1993, pp. 179-194.

[29] M. Studený, P. Boček: CI-models arising among 4 random variables, in Proceedings of WUPES'94, September 11-15, 1994, Třešť, Czech Republic, pp. 268-282.

[30] M. Studený: Description of structures of conditional stochastic independence by means of faces and imsets (a series of 3 papers), International Journal of General Systems 23 (1994/1995), n. 2-4, pp. 123-137, 201-219, 323-341.

[31] M. Studený, R. R. Bouckaert: On chain graph models for description of conditional independence structures, Annals of Statistics 26 (1998), n. 4, pp. 1434-1495.

[32] M. Studený, R. R. Bouckaert, T. Kočka: Extreme supermodular set functions over five variables, research report n. 1977, Institute of Information Theory and Automation, Prague, January 2000.

[33] M. Studený, On mathematical description of probabilistic conditional independence structures, a survey monograph in preparation.

[34] Y. Xiang, S. K. M. Wong, N. Cercone: Critical remarks on single link search in learning belief networks, in Uncertainty in Artificial Intelligence 12 (E. Horvitz, F. Jensen eds.), Morgan Kaufmann 1996, pp. 564-571.