Proceedings of the

8th WORKSHOP ON

UNCERTAINTY PROCESSING

LIBLICE

September 19 – 23, 2009

Programme Committee:

Francesc Esteva, *Spain*
Lluis Godo, *Spain*
Petr Hájek, *Czech Republic*
Radim Jiroušek, *Czech Republic*
Gernot D. Kleiter, *Austria*
Romano Scozzafava, *Italy*

Proceedings Editors:

Tomáš Kroupa, *Czech Republic*
Jiřina Vejnarová, *Czech Republic*

Organizing Chairs:

Radim Jiroušek *Czech Republic*
Gernot D. Kleiter *Austria*

ISBN 978-80-245-1543-4

8th WORKSHOP ON
UNCERTAINTY PROCESSING

organized by

Institute of Information Theory and Automation
of the Academy of Sciences of the Czech Republic

and

Faculty of Management, University of Economics

Liblice, September 19-23, 2009

with the financial support from

EUROPEAN SCIENCE FOUNDATION

EUROCORES Programme
European Collaborative Research

Foreword

The proceedings you are holding in your hands contain contributions presented at the 8th Workshop on Uncertainty Processing, WUPES'09, held in Liblice (Czech Republic), September 19–23, 2009. These workshops, the first of which was organized back in 1988, aim to foster creative intellectual activities and the exchange of ideas in an informal atmosphere. To ensure the intended goal, the tradition is to limit the number of participants and, in order to encourage discussions based on more detailed knowledge of the presented ideas, publish the proceedings containing all presented contributions. It is also a tradition that after the meeting selected papers from the workshop are published as a special issue of a renowned international journal. This year we have made a preliminary agreement with the International Journal of Approximate Reasoning.

In 2009 the workshop took place in an 18th century château designed and built for Arnošt Pacht of Rájov by Giovanni Battista Alliprandi, a famous and much sought-after Italian architect, and the builder of the most distinguished Baroque structures in Bohemia. A recent restoration of the château, which was co-financed by the European Union, has transformed it into a contemporary conference center equipped with all the necessary technical facilities. The unique union between the Baroque architecture with state-of-the-art conference premises offers an exceptional environment contributing to the informal and fruitful scientific atmosphere of the workshop.

WUPES'09 was organized jointly by the Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic and by the Faculty of Management at the University of Economics. This year's workshop is particularly special as it also includes a *LogICCC Working Day*, made possible thanks to the European Collaborative Research (EUROCORES) Programme LogICCC. This not only enables the meeting of researchers from different LogICCC Collaborative Research Projects, but it also provides an opportunity for familiarization of the representatives of the world-wide research community with results achieved within this joint European research activity. We want to stress that we greatly appreciate the LogICCC travel grants awarded to 10 researchers, covering complete travel expenses and conference fees for WUPES'09. We also want to acknowledge financial support from grant number 1M0572 awarded by MŠMT ČR.

It is quite natural that such a scientific meeting could not materialize if it were not for a hard work of a number of our friends and colleagues who did their best to guarantee its success. Though we cannot name all of them we want to express our special gratitude to all the members of the Program and Organizing Committees, the staff in the Conference center Liblice, and in the financial departments of the organizing institutions. Special thanks are due to WUPES web designer Václav Kratochvíl and the editors of these proceedings, Jiřina Vejnarová and Tomáš Kroupa, and to its cover designer Jiří Přibil.

Gernot D. Kleiter Radim Jiroušek

Table of Contents

Exploiting Unconditional Independencies in Semigraphoid Closure Computation

Marco Baioletti
Dip. Matematica e Informatica,
Università di Perugia, Italy,
baioletti@dipmat.unipg.it

Giuseppe Busanello **Barbara Vantaggi**
Dip. Metodi e Modelli Matematici,
Università "La Sapienza" Roma, Italy,
{busanello, vantaggi}@dmmm.uniroma1.it

Abstract

We deal with the problem of computing efficiently the closure with respect to semigraphoid and graphoid properties of a given set of independencies compatible with a probability. In particular, we improve the procedure by using properly unconditional independence statements.

1 Introduction

Conditional independence structures arise in different frameworks, in particular in probability. Given a coherent probability P and a set J of conditional independence statements, a related problem is to check whether the set J is compatible with P and then, to find all the set of independencies deducible from J. Then, we need to compute the closure of given set J of independencies, compatible with a (coherent conditional) probability.

We recall that, under the classical definition of independence, the independence model $\mathcal{M}$ induced by any probability P is a semigraphoid and when the probability is strictly positive, $\mathcal{M}$ is a graphoid structure. Also under the definition given in [2], for coherent conditional probability assessment graphoid properties have been tested [8].

The computation of the closure, with respect to graphoid properties (as well as with respect to semigraphoid ones) is infeasible since its size is exponentially larger than the size of the given set J of independence statements (see [6, 7]). Then, it is necessarily to find suitable reduced set of independence statements (obviously included in the closure of J with respect to graphoids), which is as smallest as possible and it represents the same independence structure. From this reduced set all the relations in the closure should be easily deducible, then it can be considered a basis for the closure. This topic by considering essentially semigraphoid structures has already been successfully solved by Studený in [6, 7]. While the case of graphoid structures has been studied in [1], where in particular we recall the basis for the closure of J, fast closure, and we give the algorithm FC1$(\cdot)$ to compute the fast closure. In particular, FC1$(\cdot)$ uses a characterization given in [1] of the closure of a pair of triples (see [1]). By using the characterization of the closure, with respect to semigraphoids, of a pair of

triples given in [6], we are able to introduce an algorithm $\mathrm{FC1}_s(\cdot)$ similar to FC1.

The aim of this work is to find in an efficient way the closure of J under semigraphoid and graphoid properties by improving the algorithms $\mathrm{FC1}(\cdot)$ and $\mathrm{FC1}_s(\cdot)$ which allows to save space and time.

The paper is organized as follows: in Section 2 the main notions on semigraphoid and graphoid are recalled. In Section 3 the definition of generalized contraction, generalized intersection and related properties are recalled and $\mathrm{FC1}_s(\cdot)$ is given. In Section 4 a new algorithm $\mathrm{FC}(\cdot)$ to computes the closure with respect to both semigraphoid and graphoid is introduced.

2 Graphoid

Let $\tilde{S} = \{Y_1, \ldots, Y_n\}$ be a finite not empty set of variables and $S = \{1, \ldots, n\}$ the set of indices associated to $\tilde{S}$. Given a (coherent) probability P, a conditional independence statement $Y_A \perp\!\!\!\perp Y_B | Y_C$ (compatible with P), where A, B, C are disjoint subsets of S, is simply denoted by the ordered triple (A, B, C). We denote with $S^{(3)}$ the set of all ordered triples (A, B, C) of disjoint subsets of S, such that A and B are not empty.

We recall that a conditional independence model $\mathcal{I}$, related to P, is a subset of $S^{(3)}$. The properties of such models depend obviously on the independence notion taken into account (see [3] for models under the classical definition and [8] under cs-independence introduced in [2]). An independence model arising from the classical independence notion is closed under semigraphoid properties, that are the following ones:

G1 if $(A, B, C) \in \mathcal{I}$, then $(B, A, C) \in \mathcal{I}$ (Symmetry);

G2 if $(A, B, C) \in \mathcal{I}$, then $(A, B', C) \in \mathcal{I}$ for any nonempty subset B' of B (Decomposition);

G3 if $(A, B_1 \cup B_2, C) \in \mathcal{I}$ with B_1 and B_2 disjoint, then $(A, B_1, C \cup B_2) \in \mathcal{I}$ (Weak Union);

G4 if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, D) \in \mathcal{I}$, then $(A, B \cup C, D) \in \mathcal{I}$ (Contraction).

If the probability is strictly positive the model is also closed under graphoid properties, it means that G1–G4 hold together with the following rule

G5 if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, B \cup D) \in \mathcal{I}$, then $(A, B \cup C, D) \in \mathcal{I}$ (Intersection).

While the model arising from cs–independence is not necessarily closed with respect to symmetry but, by reinforcing cs–independence (by requiring symmetry) the associated model is closed with respect to graphoid properties.

A semigraphoid (graphoid) is a couple $(S, \mathcal{I})$ satisfying the properties G1–G4 (G1–G5).

3 Generalized inference rules

Given a pair of triples $\theta_1, \theta_2 \in S^{(3)}$, we say that θ_2 is *generalized–included* in θ_1 (briefly g–included), in symbol $\theta_2 \sqsubseteq \theta_1$, if θ_2 can be obtained from θ_1 with a finite number of applications of G1, G2 and G3.

In [1] we prove that given $\theta_1 = (A_1, B_1, C_1)$ and $\theta_2 = (A_2, B_2, C_2)$, $\theta_1 \sqsubseteq \theta_2$ if and only if the following conditions hold

(i) $C_2 \subseteq C_1 \subseteq X_2$;

(ii) either $A_1 \subseteq A_2$ and $B_1 \subseteq B_2$ or $A_1 \subseteq B_2$ and $B_1 \subseteq A_2$;

where X_i stands for $(A_i \cup B_i \cup C_i)$, $i = 1, 2$.

Generalized inclusion is strictly related to the dominance notion introduced in [6] and denoted by $\sqsubseteq_a$ in the following. In fact, the relation between $\sqsubseteq$ and $\sqsubseteq_a$ is simple: $\theta' \sqsubseteq \theta$ if and only if either $\theta' \sqsubseteq_a \theta$ or $\theta' \sqsubseteq_a \theta^T$, where θ^T is the transpose of θ ($\theta = (A, B, C)$ then $\theta^T = (A^T, B^T, C^T)$).

The g–inclusion between triples is extended to the case of sets of triples.

Definition 1 *Let H, J be subsets of $S^{(3)}$. J is a covering of H (in symbol $H \sqsubseteq J$) if and only if for any triple $\theta \in H$ there exists a triple $\theta' \in J$ such that $\theta \sqsubseteq \theta'$.*

Our target in [1] (as that in [6] for semigraphoids), is to find an efficient method to compute a reduced (with respect to g–inclusion $\sqsubseteq$) set J^* included in the closure $\bar{J}$ and having the same information of $\bar{J}$; this means that for any triple $\theta \in \bar{J}$ there exists a triple $\theta' \in J^*$ such that $\theta \sqsubseteq \theta'$.

Given $\theta_1 = (A_1, B_1, C_1), \theta_2 = (A_2, B_2, C_2) \in S^{(3)}$, let

$$W_C(\theta_1, \theta_2) = \{\tau : \theta_1', \theta_2' \vdash_{G4} \tau, \text{ with } \theta_1' \sqsubseteq_a \theta_1, \theta_2' \sqsubseteq_a \theta_2\}.$$

In [1] (see also [6]) we give a characterization of $WC(\theta_1, \theta_2)$ and we prove that if $W_C(\theta_1, \theta_2)$ is not empty then

$$gc(\theta_1, \theta_2) = (A_1 \cap A_2, (B_1 \setminus C_2) \cup (B_2 \cap X_1), C_2 \cup (A_2 \cap C_1))$$

is in $W_C(\theta_1, \theta_2)$ and dominates any triple belonging to $W_C(\theta_1, \theta_2)$.

We denote with $GC(\theta_1, \theta_2)$ the set formed by the possible (i.e. belonging to $S^{(3)}$) triples among $gc(\theta_1, \theta_2)$, $gc(\theta_1, \theta_2^T)$, $gc(\theta_1^T, \theta_2)$ and $gc(\theta_1^T, \theta_2^T)$. Obviously, $GC(\theta_1, \theta_2)$ is in general different from $GC(\theta_2, \theta_1)$.

A similar result [1, 6] holds for

$$W_I(\theta_1, \theta_2) = \{\tau : \theta_1', \theta_2' \vdash_{G5} \tau, \text{ with } \theta_1' \sqsubseteq_a \theta_1, \theta_2' \sqsubseteq_a \theta_2\}.$$

In particular, if $W_I(\theta_1, \theta_2)$ is not empty, then

$$gi(\theta_1, \theta_2) = (A_1 \cap A_2, (B_1 \cap X_2) \cup (B_2 \cap X_1), (C_1 \cap A_2) \cup (C_2 \cap A_1) \cup (C_2 \cap C_1))$$

is in $W_I(\theta_1, \theta_2)$ and dominates any triple belonging to $W_I(\theta_1, \theta_2)$. The set $GI(\theta_1, \theta_2)$ is formed by the possible (i.e. belonging to $S^{(3)}$) triples among $gi(\theta_1, \theta_2)$, $gi(\theta_1, \theta_2^T)$, $gi(\theta_1^T, \theta_2)$ and $gi(\theta_1^T, \theta_2^T)$.
Then, $GI(\theta_1, \theta_2) = GI(\theta_2, \theta_1)$.

The previous sets $GC(\cdot, \cdot)$ and $GI(\cdot, \cdot)$ are necessary for defining the two new inference rules

G4* "generalized contraction": from θ_1, θ_2 deduce any triple $\tau \in GC(\theta_1, \theta_2)$;

G5* "generalized intersection": from θ_1, θ_2 deduce any triple $\tau \in GI(\theta_1, \theta_2)$;

which generalize the two classical inference rules.

Given a set J of triples in $S^{(3)}$, we denote with $J^* = \{\tau : J \vdash^*_G \tau\}$[1] the closure of J with respect to G4* and G5*. In [1] we prove that $J^* \subseteq \bar{J}$ and $\bar{J} \sqsubseteq J^*$.

Actually, J^* contains some "redundant" triples, that means that are g–included in the other ones.

Starting from a set $J \subseteq S^{(3)}$, in order to reduce as much as possible the cardinality of $\bar{J}$ without losing information, we define the "maximal"(with respect to g–inclusion) triple set

$$J_{/\sqsubseteq} = \{\tau \in J : \nexists \bar{\tau} \in J \text{ with } \bar{\tau} \neq \tau, \tau^T \text{ such that } \tau \sqsubseteq \bar{\tau}\}. \tag{1}$$

Obviously, $J_{/\sqsubseteq} \subseteq J$.

Definition 2 *A subset J of $S^{(3)}$ is said maximal if $J = J_{/\sqsubseteq}$.*

Moreover, by using $\bar{J}_{/\sqsubseteq}$ instead of $\bar{J}$ there is no loss of information. In fact, $\bar{J} \sqsubseteq \bar{J}_{/\sqsubseteq}$. Then, given a set J of triples in $S^{(3)}$, we compute J^* and then we cut redundant triples by taking its "maximal" triples, i.e. $J^*_{/\sqsubseteq}$. We call the set $J^*_{/\sqsubseteq}$ "fast closure" and we denote it, for simplicity, with J_*.

Proposition 1 *Let J, H be two maximal sets of $S^{(3)}$, then $H \sqsubseteq J$ and $J \sqsubseteq H$ if and only if for any $\theta \in H$ either $\theta \in J$ or $\theta^T \in J$ and for any $\tau \in J$ either $\tau \in H$ or $\tau^T \in H$.*

If $H \sqsubseteq J$ and $J \sqsubseteq H$, then H and J are said to have *equivalent information* and it is denoted as $H \equiv J$ (or $J \equiv H$).

In [1] we prove that $\bar{J}_{/\sqsubseteq} \equiv J_*$, moreover, we look for a unique inferential rule and provide an algorithm for computing J_*.

We recall first of all that the fast closure $\{\theta_1, \theta_2\}_*$ of a couple $\theta_1, \theta_2 \in S^{(3)}$ is composed by a maximum of nine extra triples, no matter how many variables occur in θ_1 and θ_2.

In particular, any pair of triples (θ_1, θ_2) can be written, in a general form, as $\theta_1 = (E_{(1,1)}, E_{(1,2)}, E_{(1,3)})$ and $\theta_2 = (E_{(2,1)}, E_{(2,2)}, E_{(2,3)})$. Each triple of the fast closure of (θ_1, θ_2) is g–included in the set of possible (i.e. belonging to $S^{(3)}$) triples

$$K(\theta_1, \theta_2) = \{\theta_1, \theta_2, \nu(\theta_1, \theta_2), \hat{\theta}_{(i,j,k)}(\theta_1, \theta_2) \, : \, i, j, k \in \{1, 2\}\} \tag{2}$$

where

$$\begin{aligned} \hat{\theta}_{(i,j,k)}(\theta_1, \theta_2) &= \Big(E_{(i,j)} \cap E_{(3-i,k)}, E_{(i,3-j)} \cup (E_{(3-i,3-k)} \cap X_i), C\Big); \\ \nu(\theta_1, \theta_2) &= \Big((E_{(1,1)} \cap E_{(2,2)}) \cup (E_{(1,2)} \cap E_{(2,1)}), (E_{(1,1)} \cap E_{(2,1)}) \cup \\ &\quad \cup (E_{(1,2)} \cap E_{(2,2)}), E_{(1,3)} \cup E_{(2,3)}\Big) \end{aligned}$$

[1] $J \vdash^*_G \tau$ means that τ is obtained by applying a finite number of times the rules G4* and G5*.

with $C = (E_{(1,3)} \cap E_{(2,3)}) \cup (E_{(i,3)} \cap E_{(3-i,k)}) \cup (E_{(3-i,3)} \cap E_{(i,j)})$.

Moreover, in [1] it is proved that $K(\theta_1, \theta_2)_{/\sqsubseteq} \equiv \{\theta_1, \theta_2\}_*$.

Note that in general $K(\theta_1, \theta_2)$ may not coincide with $\{\theta_1, \theta_2\}_*$ because it could contain some redundant triple or the transpose triple of one belonging to $\{\theta_1, \theta_2\}_*$. Therefore, the set $K(\theta_1, \theta_2)$ allows to compute a maximal covering set of $\{\theta_1, \theta_2\}_*$ having an equivalent information. All this computation requires a constant number of steps with respect to the size of θ_1, θ_2. Then, the function FC1$(\cdot)$ given in [1] is based on the inference rule:

U : from θ_1, θ_2 deduce any triple $\tau \in \{\theta_1, \theta_2\}_*$.

For semigraphoids characterization, similar to (2) for a pair of triples θ_1, θ_2, is in [6] and so it is possible to define the set

$$K_s(\theta_1, \theta_2) = \{\theta_1, \theta_2, \gamma_{(i,j,k)}(\theta_1, \theta_2), \delta_{(i,j,k)}(\theta_1, \theta_2), \nu(\theta_1, \theta_2) : i, j, k \in \{1, 2\}\},$$

where

$$\begin{aligned} \gamma_{(i,j,k)}(\theta_1, \theta_2) &= (E_{(i,j)} \cap E_{(3-i,k)}, (E_{(3-i,3-k)} \setminus E_{(i,3)}) \cup (E_{(i,3-j)} \cap X_{3-i}), L); \\ \delta_{(i,j,k)}(\theta_1, \theta_2) &= (E_{(i,j)} \cap E_{(3-i,k)}, E_{(i,3-i)} \cup (E_{(i,j)} \cap E_{(3-i,3-k)}), L); \end{aligned}$$

with $L = E_{(i,3)} \cup (E_{(i,j)} \cap E_{(3-i,3)})$.

Given a set J of triple, by denoting with $semi(J)$ the closure of J with respect to semigraphoid and by $semi(J)_*$ the related maximal set, it follows [6]

$$semi(\{\theta_1, \theta_2\})_* \equiv K_s(\theta_1, \theta_2)_{/\sqsubseteq}.$$

Also for semigraphoids it is possible to define a new inference rule

U_s : from θ_1, θ_2 deduce any triple $\tau \in semi(\{\theta_1, \theta_2\})_*$.

Moreover, by using U_s instead of U, it is simple to obtain a new algorithm FC1$_s(\cdot)$ by modifying FC1$(\cdot)$.

In the following, given a pair of triple θ_1, θ_2, $N(\theta_1, \theta_2)$ is equal to $\{\theta_1, \theta_2\}_*$ by considering the closure of $\{\theta_1, \theta_2\}$ with respect to graphoids; otherwise $N(\theta_1, \theta_2)$ is equal to $semi(\{\theta_1, \theta_2\})_*$ by considering the closure with respect to semigraphoids. Analogously, FC$_*(\cdot)$ is equal to FC1$(\cdot)$ or FC1$_s(\cdot)$ by taking the closure with respect to graphoids or semigraphoids, respectively.

In general, given set $J \subseteq S^{(3)}$, the function FC$_*(\cdot)$, to compute the closure, generates a sequence of sets

$$J_h(J) = \begin{cases} J, & h = 0; \\ (J_{h-1}(J) \cup \{\theta : \theta \in N(\theta_1, \theta_2), \theta_1, \theta_2 \in J_{h-1}(J)\})_{/\sqsubseteq}, & h > 0. \end{cases} \tag{3}$$

In particular, FC$_*(J)$=$J_{k-1}(J)$ if k is the smaller number such that $J_k(J) = J_{k-1}(J)$.

4 Closure by projection

The aim of this section is to introduce an improvement of the algorithm FC$_*(\cdot)$. In fact, given a set $J \subseteq S^{(3)}$, when unconditional statements as $\theta = (A, B, \emptyset)$,

with $A \cup B = S$, are present instead of computing $\mathrm{FC}_*(J)$, it is possible to compute the sets J_*^A, J_*^B having, with θ, the same information of $\mathrm{FC}_*(J)$, such that J_*^A and J_*^B are closed with respect to the unique inference rule U in the case of graphoids, and U_s for semigraphoids; moreover $J_*^A \cup J_*^B \cup \{\theta\}$ has cardinality not greater than $\mathrm{FC}_*(J)$. To achieve this goal we need to recall some preliminary results given in [1] based on the following notion of projection of a triple.

Definition 3 *Given $\theta = (A, B, C)$ and $Y \subseteq (A \cup B \cup C)$, if $(A \cap Y) \neq \emptyset$ and $(B \cap Y) \neq \emptyset$, then*

$$\pi_Y(\theta) = (A \cap Y, B \cap Y, C \cap Y)$$

is said the projection of θ on Y.

Now, it is straightforward to prove that

Lemma 1 *Given $\theta_1 = (A_1, B_1, C_1)$, $\theta_2 = (A_2, B_2, C_2)$ and $Y \subseteq X_1$ with $(A_1 \cap Y) \neq \emptyset$, $(B_1 \cap Y) \neq \emptyset$. If $\theta_1 \sqsubseteq_a \theta_2$, then $\pi_Y(\theta_1) \sqsubseteq_a \pi_Y(\theta_2)$.*

Given a pair of triples $\theta_1 = (A_1, B_1, C_1)$, $\theta_2 = (A_2, B_2, C_2)$ with $C_1 \subseteq X_2$ and $C_2 \subseteq X_1$, let Y be a subset of $X_1 \cap X_2$ such that the projections $\pi_Y(\theta_1)$ and $\pi_Y(\theta_2)$ are defined, we prove in [1] that

- if $gc(\pi_Y(\theta_1), \pi_Y(\theta_2)) \neq \perp$, then $gc(\theta_1, \theta_2) \neq \perp$ and $gc(\pi_Y(\theta_1), \pi_Y(\theta_2)) = \pi_Y(gc(\theta_1, \theta_2))$;
- if $gi(\pi_Y(\theta_1), \pi_Y(\theta_2)) \neq \perp$, then $gi(\theta_1, \theta_2) \neq \perp$, and $gi(\pi_Y(\theta_1), \pi_Y(\theta_2)) = \pi_Y(gi(\theta_1, \theta_2))$.

We prove in [1] also how projcetion works with respect to $K(\theta_1, \theta_2)$.

Lemma 2 *Given $\theta_1 = (A_1, B_1, C_1)$, $\theta_2 = (A_2, B_2, C_2)$ with $C_1 \subseteq X_2$ and $C_2 \subseteq X_1$, let Y be a subset of $X_1 \cap X_2$ such that the projections $\pi_Y(\theta_1)$ and $\pi_Y(\theta_2)$ are defined.*
If $\hat{\theta}_{(i,j,k)}(\pi_Y(\theta_1), \pi_Y(\theta_2)) \neq \perp$, then $\hat{\theta}_{(i,j,k)}(\theta_1, \theta_2) \neq \perp$ and

$$\hat{\theta}_{(i,j,k)}(\pi_Y(\theta_1), \pi_Y(\theta_2)) = \pi_Y(\hat{\theta}_{(i,j,k)}(\theta_1, \theta_2)).$$

If $\nu(\pi_Y(\theta_1), \pi_Y(\theta_2)) \neq \perp$, then $\nu(\theta_1, \theta_2) \neq \perp$ and

$$\nu(\pi_Y(\theta_1), \pi_Y(\theta_2)) = \pi_Y(\nu(\theta_1, \theta_2)).$$

It is simple to extend to $K_s(\theta_1, \theta_2)$ the result of the previous lemma.

Lemma 3 *Given $\theta_1 = (A_1, B_1, C_1)$, $\theta_2 = (A_2, B_2, C_2)$ with $C_1 \subseteq X_2$ and $C_2 \subseteq X_1$, let Y be a subset of $X_1 \cap X_2$ such that the projections $\pi_Y(\theta_1)$ and $\pi_Y(\theta_2)$ are defined.*
If $\gamma_{(i,j,k)}(\pi_Y(\theta_1), \pi_Y(\theta_2)) \neq \perp$, then $\gamma_{(i,j,k)}(\theta_1, \theta_2) \neq \perp$ and

$$\gamma_{(i,j,k)}(\pi_Y(\theta_1), \pi_Y(\theta_2)) = \pi_Y(\gamma_{(i,j,k)}(\theta_1, \theta_2)).$$

If $\delta(\pi_Y(\theta_1), \pi_Y(\theta_2)) \neq \perp$, then $\delta_{(i,j,k)}(\theta_1, \theta_2) \neq \perp$ and

$$\delta_{(i,j,k)}(\pi_Y(\theta_1), \pi_Y(\theta_2)) = \pi_Y(\delta_{(i,j,k)}(\theta_1, \theta_2)).$$

Proof: The proof is a straightforward consequence from the definition of the function $\gamma_{(i,j,k)}(\cdot,\cdot)$ and $\delta_{(i,j,k)}(\cdot,\cdot)$ and the properties of π_Y. □

In the following, we study an extension of the previous properties related to projection on sets of triples.

Definition 4 *Let J be a subset of $S^{(3)}$ and $Y \subseteq S$ then*

$$\pi_Y(J) = \{\pi_Y(\theta) : \theta \in J\}.$$

By definition of projection it follows $\pi_Y(\theta^T) = \pi_Y(\theta)^T$.

The next propositions shows that $\pi_Y(K(\theta_1,\theta_2)) = K(\pi_Y(\theta_1),\pi_Y(\theta_2))$ and $\pi_Y(K_s(\theta_1,\theta_2)) = K_s(\pi_Y(\theta_1),\pi_Y(\theta_2))$.

Proposition 2 *Let θ_1,θ_2 be triples of $S^{(3)}$, then for any $Y \subseteq S$*

$$\pi_Y(K(\theta_1,\theta_2)) = K(\pi_Y(\theta_1),\pi_Y(\theta_2)).$$

Proof: It is simple to see that $K(\pi_Y(\theta_1),\pi_Y(\theta_2)) =$
$\{\pi_Y(\theta_1),\pi_Y(\theta_2),\nu(\pi_Y(\theta_1),\pi_Y(\theta_2)),\hat{\theta}_{(i,j,k)}(\pi_Y(\theta_1),\pi_Y(\theta_2)) : i,j,k \in \{1,2\}\} =$
$= \pi_Y(K(\theta_1,\theta_2))$,
where the first equality is given by Definition 4 and the second one by Lemma 2. □

Proposition 3 *Let θ_1,θ_2 be triples of $S^{(3)}$, then for any $Y \subseteq S$*

$$\pi_Y(K_s(\theta_1,\theta_2)) = K_s(\pi_Y(\theta_1),\pi_Y(\theta_2)).$$

Proof: It is simple to see that $K_s(\pi_Y(\theta_1),\pi_Y(\theta_2)) =$
$= \{\pi_Y(\theta_1),\pi_Y(\theta_2),\nu(\pi_Y(\theta_1),\pi_Y(\theta_2)),\gamma_{(i,j,k)}(\pi_Y(\theta_1),\pi_Y(\theta_2),\delta_{(i,j,k)}(\pi_Y(\theta_1),\pi_Y(\theta_2)) :$
$i,j,k \in \{1,2\}\} = \pi_Y(K(\theta_1,\theta_2))$,
where the first equality is given by Definition 4 and the second one by Lemma 3. □

The following result shows how the projection works in the case of sets g–included.

Proposition 4 *Let H, J be subsets of $S^{(3)}$ such that $H \sqsubseteq J$, then for any $Y \subseteq S$*

$$\pi_Y(H) \sqsubseteq \pi_Y(J).$$

Proof: The proof it is trivial by Definition 1 and Lemma 1. □

From the previous result, we can prove the following one related to maximal sets.

Proposition 5 *Let J be a subset of $S^{(3)}$, then for any $Y \subseteq S$*

$$\pi_Y(J_{/\sqsubseteq}) \equiv \pi_Y(J)_{/\sqsubseteq}.$$

Proof: By definition of maximal set (see (1)) one has

- for any $\tau \in J_{/\sqsubseteq}$ either $\pi_Y(\tau) \in J$ or $\pi_Y(\tau)^T \in \pi_Y(J)$. Therefore, either $\tau \in \pi_Y(J_{/\sqsubseteq})$ or there exists $\tau' \in \pi_Y(J_{/\sqsubseteq})$ such that $\tau \sqsubseteq \tau'$. Then, $\pi_Y(J_{/\sqsubseteq}) \sqsubseteq \pi_Y(J)_{/\sqsubseteq}$.
- There exists $\tau^* \in \pi_Y(J)_{/\sqsubseteq}$ such that $\tau^* = \pi_Y(\tau)$ or $\tau^* = \pi_Y(\tau^T)$ with τ or $\tau^T \in J$. Therefore, $\pi(\tau) \in \pi_Y(J_{/\sqsubseteq})$ or there exists $\tau' \in J_{/\sqsubseteq}$ such that $\tau \sqsubseteq \tau'$ and $\pi_Y(\tau) \sqsubseteq \pi_Y(\tau')$. Then, $\pi_Y(J)_{/\sqsubseteq} \sqsubseteq \pi_Y(J_{/\sqsubseteq})$. □

Given a set J, the projection of its maximal set is not uniquely defined.

Example 1 *Given the set* $J = \{(\{1,2\},\{3,4\},\{5\}),(\{1,3\},\{2,4\},\{5\})\} = J_{/\sqsubseteq}$ *and let* $Y = \{2,3,5\}$ *be a subset of* $X = \{1,2,3,4,5\}$. *Then, the set* $\pi_Y(J_{/\sqsubseteq})$ *can be equal to either* $\{(\{2\},\{3\},\{5\})$ *or* $\{(\{3\},\{2\},\{5\})$.

Remark 1 *It is possible to observe that*

- $K(\pi_Y(\theta_1),\pi_Y(\theta_2))_{/\sqsubseteq} \equiv \pi_Y(K(\theta_1,\theta_2)_{/\sqsubseteq}) \equiv \pi_Y(\{\theta_1,\theta_2\}_*)$;
- $K_s(\pi_Y(\theta_1),\pi_Y(\theta_2))_{/\sqsubseteq} \equiv \pi_Y(K_s(\theta_1,\theta_2)_{/\sqsubseteq}) \equiv \pi_Y(semi(\theta_1,\theta_2)_*)$;
- $\pi_Y(N(\theta_1,\theta_2)) \equiv N(\pi_Y(\theta_1),\pi_Y(\theta_2))$.

Therefore, it is easy to verify that $\pi_Y(\{\theta_1,\theta_2\}_*) \equiv \{\pi_Y(\theta_1),\pi_Y(\theta_2)\}_*$.

Theorem 4 *Let* J *be a subset of* $S^{(3)}$, *then for any* $Y \subseteq S$ *and any* $J_h(J)$ *as in (3) with* $h \in \{0,...,k\}$ *the following conditions hold*

1. $FC_*(J) \equiv FC_*(J_h(J))$;
2. $\pi_Y(FC_*(J)) \equiv \pi_Y(FC_*(J_h(J)))$.

Proof: By definition of FC_* it is simple to observe that
$\mathrm{FC}_*(J) = \mathrm{FC}_*(J_0(J)) \equiv \mathrm{FC}_*(J_1(J)) \equiv ... \equiv \mathrm{FC}_*(J_{k-1}(J)) \equiv \mathrm{FC1}_*(J_k(J))$.
From the previous equalities trivially follows that for $h \in \{0,...,k\}$
$\pi_Y(\mathrm{FC}_*(J)) \equiv \pi_Y(\mathrm{FC}_*(J_h(J)))$. □

Now, given a set J, we show that the projection of $\mathrm{FC}_*(J)$ is equal to apply $\mathrm{FC}_*(\cdot)$ to the projection of J.

Theorem 5 *Let* J *be a subset of* $S^{(3)}$, *then for any* $Y \subseteq S$

$$\pi_Y(FC_*(J)) \equiv FC_*(\pi_Y(J)).$$

Proof: In the following with $J_0(\pi_Y(J)), J_1(\pi_Y(J)), ...$ are denoted the sets generated by FC1 to compute the fast closure of $\pi_Y(J)$. By induction we have

$$J_0(\pi_Y(J)) = \pi_Y(J) = \{\pi_Y(\theta) : \theta \in J\} = \pi_Y(J_0(J)),$$

and by supposing that $J_{h-1}(\pi_Y(J)) \equiv \pi_Y(J_{h-1}(\pi_Y(J)))$ we need to prove that $J_h(\pi_Y(J)) \equiv \pi_Y(J_h(\pi_Y(J)))$

$$J_h(\pi_Y(J)) = (J_{h-1}(\pi_Y(J)) \cup \{\theta : \theta \in \{\theta_1, \theta_2\}_*, \theta_1, \theta_2 \in J_{h-1}(\pi_Y(J))\})_{/\sqsubseteq} \equiv$$
$$\equiv (\pi_Y(J_{h-1}(J)) \cup \{\theta : \theta \in \{\theta_1, \theta_2\}_*, \theta_1, \theta_2 \in \pi_Y(J_{h-1}(J))\})_{/\sqsubseteq} \equiv$$
$$\equiv (\pi_Y(J_{h-1}(J)) \cup \{\pi_Y(\theta) : \theta \in \{\theta_1, \theta_2\}_*, \theta_1, \theta_2 \in J_{h-1}(J)\})_{/\sqsubseteq} \equiv \pi_Y(J_h(J))$$

with $h > 0$.

Moreover, $\mathrm{FC}_*(J) = J_{k-1}(J)$ whether $\mathrm{FC}_*(J)$ stops after k cycles, so that $\pi_Y(\mathrm{FC}_*(J)) = \pi_Y(J_{k-1}(J)) \equiv J_{k-1}(\pi_Y(J)) = \mathrm{FC}_*(J_{k-1}(\pi_Y(J)))$ that is by Theorem 4 equal to $\mathrm{FC}_*(J_0(\pi_Y(J))) = \mathrm{FC}_*(\pi_Y(J))$. □

From Theorem 4 and 5 it follows

Corollary 6 *Let J be a subset of $S^{(3)}$, then for any $Y \subseteq S$*

$$\pi_Y(FC_*(J)) \equiv FC_*(\pi_Y(J_h(J))).$$

From the previous result we obtain the following one when there are unconditional independence statements.

Theorem 7 *Given J a subset of $S^{(3)}$, if there exists a triple $\theta = (A, B, \emptyset) \in J$ such that $A \cup B = S$, then the following conditions hold:*

1. *$\theta \in FC_*(J)$ or $\theta^T \in FC_*(J)$;*
2. *for any $\theta' \in FC_*(J)$ with $\theta' \neq \theta$ and $\theta'^T \neq \theta$, it follows that $\pi_A(\theta') \neq \perp$ or $\pi_B(\theta') \neq \perp$;*
3. *for any $\theta_A = (A_A, B_A, C_A) \in FC_*(\pi_A(J))$ and for any $\theta_B = (A_B, B_B, C_B) \in FC_*(\pi_B(J))$ then*
   - *$\forall \bar{\theta} \in \{(A_A, B_A \cup B, C_A), (A_A \cup B, B_A, C_A), (A_B, B_B \cup A, C_B), (A_B \cup A, B_B, C_B), (A_A \cup A_B, B_A \cup B_B, C_A \cup C_B), (A_A \cup B_B, B_A \cup A_B, C_A \cup C_B)\}$ it follows that $\bar{\theta} \in FC_*(J)$ or $\bar{\theta}^T \in FC_*(J)$.*

Proof: 1. It is trivial to prove that $\theta \in \mathrm{FC}_*(J)$ or $\theta^T \in \mathrm{FC}_*(J)$. In fact, if there exists $\theta' = (A', B', C') \in \mathrm{FC}_*(J)$ such that $\theta \sqsubseteq \theta'$ then $C' = \emptyset$ with $A' \cup B' = S$. Therefore, one as one of the following situations: $A = A'$ and $B = B'$ or $A = B'$ and $B = A'$.

2. If there exists a triple $\theta' = (A', B', C') \in \mathrm{FC}_*(J)$ with $\pi_A(\theta') = \perp$ and $\pi_B(\theta') = \perp$, since $A \cup B = S$ then either $A' \cap B = \emptyset$ and $B' \cap A = \emptyset$ or $A' \cap A = \emptyset$ and $B' \cap B = \emptyset$. In the first case $A' \subseteq A$ and $B' \subseteq B$, in the second one $A' \subseteq B$ and $B' \subseteq A$, then in both cases $\theta' \sqsubseteq \theta$ and this is absurd.

3. Since $\theta_A \in \mathrm{FC}_*(\pi_A(J))$ and $\theta_B \in \mathrm{FC}_*(\pi_B(J))$ there exists $\bar{\theta}_A = (\bar{A}_A, \bar{A}_B, \bar{C}_A)$, $\theta'_B = (\bar{A}_B, \bar{B}_B, \bar{C}_B) \in \mathrm{FC}_*(J)$ such that $\theta_A = \pi_A(\bar{\theta}_A)$ (or $\theta_A^T = \pi_A(\bar{\theta}_A)$) and $\theta_B = \pi_B(\bar{\theta}_B)$ (or $\theta_B^T = \pi_B(\bar{\theta}_B)$). We assume, without loss of generality, that $\theta_A = \pi_A(\bar{\theta}_A)$ and $\theta_B = \pi_B(\bar{\theta}_B)$. From the characterization given in [1] all set $W_C(\theta, \bar{\theta}_A)$, $W_C(\theta, \bar{\theta}_A^T)$, $W_C(\theta, \bar{\theta}_B)$, $W_C(\theta, \bar{\theta}_B^T)$ are not empty, so

- $\theta_{1A} = gc(\bar{\theta}_A, \theta) = (\bar{A}_A \cap A, \bar{B}_A \cup (B \cap X_{\bar{A}}), A \cap \bar{C}_A) = (A_A, B_A \cup (B \cap X_{\bar{A}}), C_A)$. Moreover $\theta'_A = gc(\theta, \theta_{1A}) = (A_A, B_A \cup B, C_A)$.
- $\theta_{2A} = gc(\bar{\theta}_A^T, \theta) = (\bar{B}_A \cap A, \bar{A}_A \cup (B \cap X_{\bar{A}}), A \cap \bar{C}_A) = (B_A, A_A \cup (B \cap X_{\bar{A}}), C_A)$. Moreover $\theta''_A = gc(\theta, \theta_{2A}) = (B_A, A_A \cup B, C_A)$.

- $\theta_{1B} = gc(\bar{\theta}_B, \theta^T) = (\bar{A}_B \cap B, \bar{B}_B \cup (A \cap X_{\bar{B}}), B \cap \bar{C}_B) = (A_B, B_B \cup (A \cap X_{\bar{B}}), C_B)$. Moreover $\theta'_B = gc(\theta^T, \theta_{1B}) = (A_B, B_B \cup A, C_B)$.
- $\theta_{2B} = gc(\bar{\theta}^T_B, \theta^T) = (\bar{B}_B \cap B, \bar{A}_B \cup (A \cap X_{\bar{B}}), B \cap \bar{C}_B) = (B_B, A_B \cup (A \cap X_{\bar{B}}), C_B)$. Moreover $\theta''_B = gc(\theta^T, \theta_{2B}) = (B_B, A_B \cup A, C_B)$.

Furthermore, also $W_C(\theta'^T_A, \theta'^T_B)$, $W_C(\theta''_A, \theta''_B)$, $W_C(\theta'^T_A, \theta''_B)$, $W_C(\theta''_A, \theta'^T_B)$ are not empty, so that

- $\theta_1 = gc(\theta'^T_A, \theta'^T_B)^T = gc(\theta''^T_A, \theta''^T_B) = (A_A \cup A_B, B_A \cup B_B, C_A \cup C_B)$;
- $\theta_2 = gc(\theta'^T_A, \theta''^T_B)^T = gc(\theta''^T_A, \theta'^T_B) = (A_A \cup B_B, A_B \cup B_A, C_A \cup C_B)$.

We show that $\theta_1 = gc(\theta'^T_A, \theta'^T_B)^T \neq \perp$.
In fact, $\theta'^T_A = (B_A \cup B, A_A, C_A)$ and $\theta'^T_B = (B_B \cup A, A_B, C_B)$ so that $gc(\theta'^T_A, \theta'^T_B) = ((B_A \cup B) \cap (B_B \cup A), (A_A \setminus C_B) \cup (A_B \cup (A_A \cup A_B \cup A_C \cup B)), C_B \cup ((A \cup B_B) \cap C_A)) = (A_B \cup B_B, A_A \cup A_B, C_A \cup C_B) = \theta^T_1$.
Analogously, it is possible to prove that $\theta_1 = gc(\theta''^T_A, \theta''^T_B)$ and $\theta_2 = gc(\theta'^T_A, \theta''^T_B)^T = gc(\theta''^T_A, \theta'^T_B)$.

Now, if there exists a triple $\theta' \in \mathrm{FC}_*(J)$ such that $\theta'_A \sqsubseteq \theta'$, $\theta'_A \neq \theta'$ and $\theta'^T_A \neq \theta'$ then $\pi_A(\theta'_A) \sqsubseteq \pi_A(\theta') \in \mathrm{FC}_*(\pi_A(J))$ by Lemma 1, moreover $\pi_A(\theta'_A) = \theta_A \neq \pi_A(\theta')$, $\pi_A(\theta'_A)^T \neq \pi_A(\theta')$ and this is not possible since $\theta_A \in \mathrm{FC}_*(\pi_A(J))$.

The other cases can be proved analogously. □

By the previous theorem the following result showing that J^{AB}_* has the same information of $\mathrm{FC}_*(J)$ can be proved.

Corollary 8 *Given a subset J of $S^{(3)}$, if there exists a triple $\theta = (A, B, \emptyset) \in J$ such that $A \cup B = S$, then*

$$FC_*(J) \sqsubseteq J^{AB}_* \qquad and \qquad J^{AB}_* \sqsubseteq FC_*(J)$$

with $J^{AB}_* = \{\theta\} \cup \{\tau : \tau \in \{(A_A, B_A \cup B, C_A), (A_A \cup B, B_A, C_A), (A_B, B_B \cup A, C_B), (A_B \cup A, B_B, C_B), (A_A \cup A_B, B_A \cup B_B, C_A \cup C_B), (A_A \cup B_B, B_A \cup A_B, C_A \cup C_B)\}, \theta_A = (A_A, B_A, C_A) \in \pi_A(FC_*(J)), \theta_B = (A_B, B_B, C_B) \in \pi_B(FC_*(J))\}$.

Proof: By condition *3.* of the Theorem 7 $J^{AB}_* \sqsubseteq \mathrm{FC}_*(J)$ trivially follows.

By Theorem 7 $\mathrm{FC}_*(J) \sqsubseteq J^{AB}_*$ easily follows, in fact, for any triple $\theta' = (A', B', C') \in \mathrm{FC}_*(J)$ we have the following situations

- $\pi_A(\theta') \neq \perp$, $\pi_B(\theta') \neq \perp$ then $\{\theta'\} \sqsubseteq J^{AB}_*$;
- $\pi_A(\theta') = (A'_A, B'_A, C'_A) \neq \perp$, $\pi_B(\theta') = \perp$ then $\{\theta'\} \sqsubseteq \{(A'_A \cup B, B'_A, C'_A), (A'_A, B'_A \cup B, C'_A)\} \sqsubseteq J^{AB}_*$;
- $\pi_A(\theta') = \perp$, $\pi_B(\theta') = (A'B, B'_B, C'_B) \neq \perp$ then $\{\theta'\} \sqsubseteq \{(A'_B \cup A, B'_B, C'_B), (A'_B, B'_B \cup A, C'_B)\} \sqsubseteq J^{AB}_*$. □

By the previous corollary and Theorem 7 the advantages of projection follow, in fact, we are able to reduce (at least the half) the number of maximal triples. For example, instead of considering $\mathrm{FC}_*(J) = \{\theta, (A_A, B_A \cup B, C_A), (A_A \cup B, B_A, C_A), (A_B, B_B \cup A, C_B), (A_B \cup A, B_B, C_B), (A_A \cup A_B, B_A \cup B_B, C_A \cup C_B), (A_A \cup B_B, B_A \cup A_B, C_A \cup C_B)\}$, by applying projection, we need to store

by starting from θ the sets $J_*^A = \{\theta = (A, B, \emptyset)\}$, $J_*^B = \{\theta_A = (A_A, B_A, C_A)\}$ and the triple $\theta_B = (A_B, B_B, C_B)$. In the other ones, where it is possible to apply the projection, we are able to reduce even more space. By the Corollary 8 we are able to define a new algorithm called simply FC to exploit the projection. The function FINDTRIPLE(J_k, S) returns either a triple $\theta = (A, B, \emptyset)$

Algorithm 1 Fast closure by projection

1: **function** FC(J, S) ▷ J is a maximal set
2: $J_0 \leftarrow J$
3: $N_0 \leftarrow J_0$
4: $k \leftarrow 0$
5: **repeat**
6: $\theta \leftarrow$ FINDTRIPLE(J_k, S)
7: ▷ either $\theta = (A, B, \emptyset)$ with $A \cup B = S$ or $\theta = \perp$
8: **if** $\theta \neq \perp$ **then**
9: $J_k^A \leftarrow$ FC($\pi_A(J_k)$, A)
10: $J_k^B \leftarrow$ FC($\pi_B(J_k)$, B)
11: **return** $< J_k^A, J_k^B, \theta >$
12: **end if**
13: $k \leftarrow k + 1$
14: $N_k := \bigcup_{\theta_1 \in J_{k-1}, \theta_2 \in N_{k-1}} N(\theta_1, \theta_2)$
15: $J_k \leftarrow$ FINDMAXIMAL($J_{k-1} \cup N_k$)
16: **until** $J_k = J_{k-1}$
17: **return** $< J_k, \emptyset, \perp >$
18: **end function**

with $A \cup B = S$, if it exists in J_k, or $\perp$ otherwise. Moreover, FINDMAXIMAL($\cdot$) computes $J_{/\sqsubseteq}$ for a given set $J \subseteq S^{(3)}$.

References

[1] Baioletti M., Busanello G., Vantaggi B. (2009), Conditional independence structure and its closure: Inferential rules and algorithms, *International Journal of Approximate Reasoning*, in press doi: 10.1016/j.ijar.2009.05.002.

[2] Coletti G., Scozzafava R. (2000), Zero probabilities in stochastical independence, *Information, Uncertainty, Fusion, Kluwer Academic Publishers, Dordrecht*, B. Bouchon- Meunier, R.R. Yager, L.A. Zadeh (Eds.), pp. 185–196.

[3] Dawid A. P. (1979), Conditional independence in statistical theory, *J. Roy. Stat. Soc. B*, 41, pp. 15–31.

[4] Lauritzen S. L. (1996), *Graphical models*. Clarendon Press, Oxford.

[5] Pearl J. (1988), *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, Los Altos, CA.

[6] Studený M. (1997), Semigraphoids and structures of probabilistic conditional independence, *Ann. Math. Artif. Intell.*, 21, pp. 71–98.

[7] Studený M. (1998), Complexity of structural models, *Proc. Prague Stochastics '98*, Prague, pp. 521–528.

[8] Vantaggi B. (2001), Conditional independence in a coherent setting, *Ann. Math. Artif. Intell.*, 32, pp. 287–313.

Coherent Conditional Possibilities in Medical Diagnosis

M. Baioletti
Dept. of Mathematics and Informatics
University of Perugia
baioletti@dipmat.unipg.it

G. Coletti
Dept. of Mathematics and Informatics
University of Perugia
coletti@dipmat.unipg.it

D. Petturiti
Dept. of Mathematics and Informatics
University of Perugia
e-mail davidepitturity@gmail.com

B. Vantaggi
Department Me.Mo.Mat.
University of Rome "La Sapienza"
vantaggi@dmmm.uniroma1.it

Abstract

The main aim of this paper is to present how coherent conditional possibilities can be useful in medical diagnosis. Given some possible diseases (that could explain an initial piece of information) and a relevant tentative possibility assessment, a doctor can have at his disposal also a data base giving rise to conditional possibilities $\Pi(E|K_i)$, where each K_i is a disease and each evidence E comes from a suitable test. Once the coherence of the whole assessment is checked, we want to suitably update the prior possibilities. Nevertheless, similarly to what discussed in a probabilistic framework (see [4]), if we do not assume that the diseases constitute a partition of the certain event Ω, we need a generalized concept of inference, consisting on an enlargement procedure of the assessment to the events $K_i|E$. Usually the result is in general *not unique* so we obtain an upper possibility (which is a possibility) and a lower possibility. We present also a sketch of the relevant algorithms and briefly discuss about their computational complexity.

1 Introduction

Recently a well founded theory of coherent (conditional) possibilities has been developed (see for instance [2, 3, 9, 10, 17]). Coherent possibility approach allows to assign possibilistic evaluations on an arbitrary set of events and then to extend it to all the set of events of interest. The degree of belief on an event of the new set turns out to be represented by an interval (defined by all the coherent extensions), rather than a single number. Starting from the conditions characterizing coherent assessments, it is possible to elaborate efficient algorithms for checking coherence and for extending the assessment to new events (see [1]). Here we present only a sketch of the relevant algorithms and briefly discuss about their computability.

Diagnosis procedures in a possibilistic framework are present in the literature (see for example [13]). In this paper, by using the theory and following the idea

exposed in [4] for a probabilistic framework, we present a procedure for handling uncertainty in the process of medical diagnosis, by using coherent conditional possibility. The proposed interactive procedure initially refers to

(i) a family of hypotheses (that is, *events* represented by suitable propositions) supplied by the physician: they correspond to possible diseases H_i ($i = 1, 2, ..., n$) which could explain a given initial piece of information referring to the specific situation (anamnesis). No structure and no simplifying and unrealistic assumption (such as mutual exclusiveness and exhaustivity) is required for this family of events;

(ii) all logical relations between these hypotheses, either already included in the knowledge base, or given by the doctor on the basis of the specific situation;

(iii) a possibility assessment on the given set of hypotheses. Clearly, this is not a complete assessment, since these events have been chosen as the most natural according to the doctor's experience: so usually they do not constitute, in general, a partition of the certain event Ω, and therefore the extension to other events of these possibility evaluations is not necessarily unique. Moreover, a doctor often assigns degrees of belief directly to *sets of hypotheses* (for example, one may suspect that one of the diseases the patient suffers from is an infectious one, but he is not able to commit any belief to particular infectious diseases);

(iv) a data base consisting of conditional events $E|K$ and their relevant possibilities $\Pi(E|K)$, where each event K is a possible disease which is in some way related to the given hypotheses H_i, while each evidence E comes from a suitable evidential test. These possibilities could have been obtained directly by means of a data-base containing possibilistic or fuzzy information or computed as an upper probability starting from a coherent probability assessment on a different suitable set of events (see for instance [12, 14, 16, 7, 8]).

Then, once this preliminary preparation has been done, the first step of our procedure consists in checking coherence of the assessment $\Pi(H_i)$. If the assessment turns out not being coherent, the doctor can be driven to a different assignment based on the relevant mathematical relations contained in the corresponding system. Another way-out is to look for suitable subfamilies of the set $\{H_1, H_2, ..., H_n\}$ for which the assignment is coherent, and then proceed by resorting to the extension theorem.

On the contrary, coherence of the possibilities $\Pi(H_i)$ allows to go on by checking now the coherence of the whole assessment including also the possibilities $\Pi(E|K)$. The whole assignment (prior possibilities and "possibilistic likelihood") can be incoherent even if the two separate assessment are coherent.

On the basis of the results obtained by means of the evidential tests, the doctor can now update the possibilities of the hypotheses H_i, i.e. he assesses the conditional possibilities $\Pi(H_i|E)$. Then he needs to check again coherence of the whole assessment including the latter and the former possibility evaluations.

When prior possibilities and possibilistic likelihood are jointly coherent, the doctor can get values representing each posterior possibility (of a disease H_i given an evidence E) by using the extension theorem. Usually the extension is not unique and we can compute upper and lower bounds for the posterior possibilities $\Pi(H_i|E)$.

2 Coherent possibility assessments

The concept of coherence, introduced by de Finetti [11] in probability theory, has a fundamental role to manage partial assessments of an uncertainty measure and its enlargement; in other words to check consistency with respect to a specific uncertainty measure and to make inference starting from this information. Some times we can obtain a (conditional) possibility assessment during the updating process, starting from a coherent probability assessment (see [7, 8]).

2.1 Coherent unconditional assessments

To illustrate the concept of coherence in the simpler case of unconditional events, consider an assessment $\Pi(E_i)$, $i = 1, 2, ..., n$, on an *arbitrary* finite family

$$\mathcal{F} = \{E_1, ..., E_n\},$$

and denote by $\mathcal{C} = \{C_1, ..., C_m\}$ the set of the atoms generated by these events (*i.e.* made up with all possible conjunctions $E_1^* \wedge E_2^* \ldots \wedge E_n^*$, different from the impossible event $\emptyset$, obtained by putting in place of each E_i^*, for $i = 1, 2, \ldots, n$, the event E_i or its contrary E_i^c). This assessment is called *coherent* (or consistent) with a possibility if the function Π can be extended from $\mathcal{F}$ to the set of atoms, in such a way that Π *is a possibility* on the algebra $\mathcal{B}$ spanned by them. This clearly amounts to the existence of at least one solution of the following system, where $\Pi'(C_r) = x_r$ with $C_r \in \mathcal{C}$,

$$S = \begin{cases} \max\limits_{C_r \subseteq E_i} x_r = \Pi(E_i) & \forall\, E_i \in \mathcal{E} \\ \max\limits_{C_r \in \mathcal{C}} x_r = 1 & \\ x_r \geq 0 & \forall\, C_r \in \mathcal{C} \end{cases} \tag{1}$$

Note that the above system S can have more than one solution. The solvability of the above system can be proved by checking some logical constraints only, as proved in the following theorem.

Theorem 1 *Let* Π *be an assessment on* $\mathcal{E} = \{E_1, ..., E_n\}$. *Suppose that* $\Pi(E_i) \leq \Pi(E_{i+1})$ *for any* $i = 1, ..., n-1$. *The following statements are equivalent:*

- Π *is coherent with a possibility;*
- $E_j \wedge \left(\bigwedge\limits_{k<j} E_k^c \right) \neq \emptyset$ *for any* $j = 2, ..., n$ *and if* $E_1 = \emptyset$, *then* $\Pi(E_1) = 0$.

  Moreover if $\Pi(E_n) < 1$, *then* $\bigwedge\limits_{i=1}^{n} E_i^c \neq \emptyset$.

Proof: Since $E_j \wedge \left(\bigwedge\limits_{k<j} E_k^c \right) \neq \emptyset$ it contains an atom C_r, and it is possible to assign to it the value $\Pi(E_j)$, for any $j = 2, ..., n$. Moreover if $E_1 \neq \emptyset$ we assign to an atom in E_1 the value $\Pi(E_1)$. Furthermore, if $\Pi(E_n) < 1$ we can give to the atom $\bigwedge\limits_{i=1}^{n} E_i^c$ the value 1. Then, this is a solution for the system (1), so the assessment is coherent with a possibility.

Vice versa if the assessment is coherent with a possibility, then the system (1) admits a solution and so for any $j = 2, ..., n$ there is an atom in $E_j \wedge \left(\bigwedge_{k<j} E_k^c\right)$, moreover if $\Pi(E_1) > 0$, then $E_1 \neq \emptyset$.

The problem of checking coherence is NP-complete, so we do not expect to find polynomial-time algorithms for it. Anyway, the above result reduces this problem to solve a sequence of at most n logic satisfiability problems and therefore an algorithm able to check coherence needs $O(n)$ calls to a SAT solver. We overcome the main computational problem consisting on generating the atoms (as required in system 1) whose number exponentially increases. Even if the algorithm is still exponential, it relies on advantages in the logic satisfiability field where SAT solvers are particularly efficient (for more details see [1]).

2.2 Extending coherent unconditional assessments

For any event $A \notin \mathcal{E}$, we denote with A_* and A^*, respectively, the maximal event logically dependent on $\mathcal{E}$ contained in A, i.e.

$$A_* = \bigvee_{C_i \subseteq A} C_i,$$

and the minimal event logically dependent on $\mathcal{E}$ containing A, i.e.

$$A^* = \bigvee_{C_i \wedge A \neq \emptyset} C_i.$$

Obviously, $A_* \subseteq A \subseteq A^*$, and if A is logically dependent on $\mathcal{E}$ (i.e. $A \in \mathcal{B}$), then $A_* = A = A^*$.

Let $\mathcal{E} = \{E_1, ..., E_n\}$ be a finite set of events and Π a coherent possibility assessment on $\mathcal{E}$, then Π can be extended as a coherent possibility on any finite $\mathcal{E}^* \supset \mathcal{E}$.

Moreover, if $\mathcal{E}^* = \mathcal{E} \cup \{A\}$, then the set of coherent values for A is a closed interval $[\pi_*(A), \pi^*(A)]$, where $\pi_*(A) = \pi_*(A_*) = \min \Pi^i(A_*)$ and $\pi^*(A) = \pi^*(A^*) = \max \Pi^i(A^*)$, and the minimum and maximum are computed over the set of all possible extensions Π^i of Π on $\mathcal{B}$.

The above result gives rise to a procedure for finding the set of coherent values for any new event A, in fact it consists in finding the extreme values $\min\left(\max_{C_r \subseteq A} x_r\right)$ and $\max\left(\max_{C_r \wedge A \neq \emptyset} x_r\right)$ under the system S.

Remark 1 *It is possible to prove (see[10]) that if we start from a coherent possibility assessment* Π *on a set* $\mathcal{E}$ *and we compute the intervals of coherence for more then one new event, then we can choose for every event the maximum of the relevant interval of coherence, obtaining again a possibility extending* Π. *In other words, contrary to probability, the "upper possibility" is a possibility.*

Remark 2 *We notice that the possible values assumed by both* $\pi_*(A)$ *and* $\pi^*(A)$ *are contained in the following set* $\{0, 1, \Pi(E_i), i = 1, ..., n\}$. *In fact,* $\pi^*(A)$ *is given by the maximum among* $\Pi(E_i)$, *for* $E_i \wedge A \neq \emptyset$, *and 1, if* $\bigwedge E_i^c \wedge A \neq \emptyset$. *Analogously,* $\pi_*(A)$ *is the maximum among* $\Pi(E_i)$, *for* $E_i \subseteq A$, *and 1, if all* $\Pi(E_i) < 1$ *and* $\bigwedge E_i^c \subseteq A$. *Therefore, even if the extension problem*

is intractable, we can provide an algorithm (see for more details [1]) able to compute lower and upper bounds with at most n calls to the coherence procedure (see [1]). More precisely, for the lower bound we start by checking coherence of the assessment $\{\Pi(A) = k, \Pi(E_i), i = 1, ..., n\}$, *with* $k \in \{0, 1, \Pi(E_i), i = 1, ..., n\}$. *In conclusion, the algorithm sketched solves the extension problem with* $O(n^2)$ *calls to a SAT solver.*

2.3 Coherent conditional assessments

We refer to conditional possibility (see [2, 3]) introduced directly as a real function defined on conditional events by means the following set of axioms :

Definition 1 *Let* $\mathcal{F} = \mathcal{B} \times \mathcal{H}$ *be a set of conditional events* $E|H$ *such that* $\mathcal{B}$ *is a Boolean algebra and* $\mathcal{H}$ *an additive set (i.e. closed with respect to finite logical sums), with* $\mathcal{H} \subset \mathcal{B}$ *and* $\emptyset \notin \mathcal{H}$. *A function* $\Pi : \mathcal{F} \rightarrow [0, 1]$ *is a conditional possibility if it satisfies the following properties:*

1. $\Pi(E|H) = \Pi(E \wedge H|H)$, *for every* $E \in \mathcal{B}$ *and* $H \in \mathcal{H}$;
2. $\Pi(\cdot|H)$ *is a possibility, for any* $H \in \mathcal{H}$;
3. $\Pi(E \wedge F|H) = \min\{\Pi(E|H), \Pi(F|E \wedge H)\}$, *for any* $H, E \wedge H \in \mathcal{H}$ *and* $E, F \in \mathcal{B}$.

Let Π be an assessment on an arbitrary finite set of conditional events $\mathcal{E}$, then Π is a coherent possibility assessment iff there exist $\mathcal{F} \supseteq \mathcal{E}$ with $\mathcal{F} = \mathcal{B} \times \mathcal{H}$, $\mathcal{B}$ Boolean algebra, $\mathcal{H} \subseteq \mathcal{B}^0$ an additive set, and a conditional possibility defined on $\mathcal{F}$, extending Π (see [9]).

To characterize coherent conditional possibility assessments we introduce the concept of agreeing class (see [9, 10]).

Definition 2 *Let* $\mathcal{B}$ *be a finite algebra and* $\mathcal{C}_0$ *be the set of atoms in* $\mathcal{B}$. *The class* $\prod = \{\Pi_0, ..., \Pi_k\}$ *of possibilities defined on* $\mathcal{B}$ *is said nested if the following conditions hold for any* $j = 1, ..., k$:

- $\Pi_j(C) = \Pi_{j-1}(C)$ *if* $C \in \mathcal{C}_j \setminus \mathcal{H}_j$ $(j > 0)$,
- $\Pi_{j-1}(C) \leq \Pi_j(C) \leq 1$ *if* $C \in \mathcal{H}_j$ $(j > 0)$,
- $\Pi_j(C) = 0$ *for all the atoms* $C \in \mathcal{C}_0 \setminus \mathcal{C}_j$,
- *for any* $C \in \mathcal{C}_0$ *there exists a (unique)* $j = 0, ..., k$ *such that* $\Pi_j(C) = 1$,

where $\mathcal{C}_j = \{C \in \mathcal{C}_{j-1} : \Pi_{j-1}(C) < 1\}$ *and*

$$\mathcal{H}_j = \{C_i \in \mathcal{C}_j : \nexists C \in \mathcal{C}_j \text{ s.t. } \Pi_{j-1}(C) > \Pi_{j-1}(C_i)\}.$$

Notice that $\mathcal{H}_j$ (with $j > 0$) is actually the set of the elements of $\mathcal{C}_j$ with the "highest" value of possibility Π_{j-1}, which potentially have possibility Π_j equal to 1 (see the second condition of Definition 2).

Definition 3 *A class* $\prod = \{\Pi_0, ..., \Pi_k\}$ *of possibilities on* $\mathcal{B}$ *is agreeing with a conditional possibility* $\Pi(\cdot|\cdot)$ *on* $\mathcal{B} \times \mathcal{H}$ *if it is nested and, for any* $E|H \in \mathcal{B} \times \mathcal{H}$, $\Pi(E|H)$ *is a solution of all the equations*

$$\Pi_\alpha(E \wedge H) = \min\{x, \Pi_\alpha(H)\} \tag{2}$$

$\alpha = 0, ..., j_* + 1$ *with* $j_* = \max\{j : \Pi_j(H) < 1\}$, *and* $\Pi(E|H)$ *is the unique solution of the above equation for* $j = j_* + 1$.

We are able now to give a characterization theorem:

Theorem 2 *Let* $\mathcal{F} = \{E_1|H_1, ..., E_n|H_n\}$ *be a finite set of conditional events,* $\mathcal{C}_0$ *and* $\mathcal{B}$ *denote the set of atoms and the algebra spanned by* $\{E_1, H_1, ..., E_n, H_n\}$, *respectively.*
For a real function $\Pi : \mathcal{F} \to [0, 1]$, *the following statements are equivalent:*

a) Π *is a coherent conditional possibility assessment on* $\mathcal{F}$*;*

b) there exists (at least) a nested class $\prod = \{\Pi_0, ..., \Pi_k\}$ *of possibilities on* $\mathcal{B}$, *such that for every* $E_i|H_i \in \mathcal{F}$ *one has that* $\Pi(E_i|H_i)$ *is a solution of all the equations*

$$\Pi_\beta(E_i \wedge H_i) = \min\{x, \Pi_\beta(H_i)\} \tag{3}$$

for every β *such that* $\Pi_\beta(H_i) \leq 1$*;*

c) there exists a sequence of compatible systems S_α $(\alpha = 0, ..., k)$, *with unknown* x_r^α,

$$S_\alpha = \begin{cases} \max\limits_{C_r \subseteq E_i \wedge H_i} x_r^\alpha = \min\{\Pi(E_i|H_i), \max\limits_{C_r \subseteq H_i} x_r^\alpha\} & \text{if } \max\limits_{C_r \subseteq H_i} \mathbf{x}_r^{\alpha-1} < 1 \\ x_r^\alpha \geq \mathbf{x}_r^{\alpha-1} & \text{if } C_r \in \mathcal{H}^\alpha \\ x_r^\alpha = \mathbf{x}_r^{\alpha-1} & \text{if } C_r \in \mathcal{C}_\alpha \setminus \mathcal{H}^\alpha \\ \max\limits_{C_r \in \mathcal{C}_\alpha} x_r = 1 & \\ x_r^\alpha \geq 0 & \forall C_r \in \mathcal{C}_\alpha \end{cases} \tag{4}$$

where $\mathbf{x}^\alpha$ *(with* r*-th component* $\mathbf{x}_r^\alpha$*) indicates a solution of the system* S_α, $\mathcal{C}_\alpha = \{C_r : \mathbf{x}_r^\alpha < 1\}$ *and* $\mathcal{H}^\alpha = \{C_r : C_r \in \mathcal{C}_\alpha, \mathbf{x}_r^{\alpha-1} = \max_{C_j \in \mathcal{C}_\alpha} \mathbf{x}_j^{\alpha-1}\}$, *moreover* $\mathbf{x}_r^{-1} = 0$ *for any* C_r *in* $\mathcal{C}_0$.

The above result implies that coherence of a given assignment Π can be proved by finding an agreeing class, i.e. checking the compatibility of the sequence of systems.

Remark 3 *We note that the solution* $\mathbf{x}^\alpha$ *must be chosen by taking for any component the maximum possible value (that is 1 or one of the values of* $\Pi(E_i|H_i)$ *present in the system). In fact in the following systems* S_β, $(\beta > \alpha)$, *the value of any* x_k^β *present in the system, must be greater or equal to* $\mathbf{x_k}^\alpha$. *So, any different choice of the value is a constrain, added by us in the systems.*

As an easy corollary of Theorem 2, it is possible to prove the following results:

Theorem 3 *Let $\mathcal{F} = \{E_1|H_1, ..., E_n|H_n\}$ be a finite set of conditional events, where the events H_i are a partition of Ω. Then any real function $\Pi : \mathcal{F} \to [0, 1]$, assigning 1 to the events $E_i|H_i$ with $H_i \subseteq E_i$ and 0 to the events $E_i|H_i$ with $H_i \wedge E_i = \emptyset$ is a coherent conditional possibility assessment.*

Theorem 4 *Let $\mathcal{H} = \{H_1, ..., H_n, \}$ be a partition of Ω and $\mathcal{F} = \{E_1|H_1, ..., E_n|H_n\}$. Then any assessment $\Pi : \mathcal{F} \cup \mathcal{H} \to [0, 1]$, assigning 1 to the events $E_i|H_i$ with $H_i \subseteq E_i$, 0 to the events $E_i|H_i$ with $H_i \wedge E_i = \emptyset$ and assigning value 1 at least to an event H_i, is a coherent conditional possibility assessment.*

Remark 4 *Obviously even in the conditional case, the problem of checking coherence is NP-complete, and also in this case we can provide an algorithm (essentially based on Remark 3) based on $O(n^3)$ logic satisfiability problems.*

This algorithm proceeds iteratively by finding a maximal possibility distribution at each step. In fact the main idea is (at each step) to give the maximum value to each E_iH_i and H_i (in particular $\Pi(E_i|H_i)$ and 1, respectively, if it is possible).

We remark that in the check of coherence, the value "1" plays under conditional possibility, the same central role that the value "0" under conditional probability (see [5]).

*In particular this procedure avoids to build all the atoms, but it verifies just the satisfiability of some compound events. Moreover, at each step, at least an equation (*min *constraint) is removed and so we are able to check coherence with $O(n^3)$ calls to a SAT solver. For more details on computational complexity, see [1].*

2.4 Extending coherent conditional assessments

In this section we study the problem of extendibility of a coherent conditional possibility. For that, we refer, given a coherent conditional possibility assessment Π on $\mathcal{E}$, to the following two points:

(i) finding all coherent extensions on $E|H$ when $E|H \in \mathcal{B} \times \mathcal{B}^o$ (i.e. $E \wedge H$ and H are logically dependent on $\mathcal{E}$);

(ii) extending this result to any conditional event $F|K$ (i.e. $F|K \notin \mathcal{B} \times \mathcal{B}^o$).

To face the case (ii) we need to consider the maximal (and minimal) conditional event in $\mathcal{B} \times \mathcal{B}^o$ contained in (containing) $F|K$, with respect to the following inclusion operation between conditional events (see, e.g. [5, 18])

$$A|H \subseteq^* B|K \Longleftrightarrow AH \subseteq BK \text{ and } B^cK \subseteq A^cH.$$

Therefore, the maximum [minimum] (with respect to $\subseteq^*$) event contained in [containing] $F|K$ is $(F \wedge K)'|\,((F \wedge K)' \vee (F^c \wedge K)'')$
[and $(F \wedge K)''|\,((F \wedge K)'' \vee (F^c \wedge K)')$] where

$(D)' = \bigvee_{C_r \subseteq D} C_r$;

$(D)'' = \bigvee_{C_r \wedge D \neq \emptyset} C_r$.

In the case (i) we evaluate all the corresponding values $\Pi(E|H)$, and then we take the minimum π_* and the maximum π^* with respect to all the possible

extensions of Π, or equivalently with respect to all the agreeing classes. This means to consider the sequence of systems S^m_α and to find the maximum and the minimum coherent values for $\Pi(E|H)$ under all the possible solutions of the system S^m_α.

This problem is equivalent to find the maximum index α^* such that the solutions $\mathbf{x}^\alpha$ of the optimal problem minimizing $\max_{C_r \subseteq E \wedge H} x^\alpha_r$ under S_α ($\alpha = 0, ..., \alpha^*$) are such that $\max_{C_r \subseteq E \wedge H} \mathbf{x}^\alpha_r = \max_{C_r \subseteq H} \mathbf{x}^\alpha_r$. The aim is, in fact, to eliminate as much as possible constraints in a way to get the extremal coherent values for $\Pi(E|H)$ by forcing the relevant equation to be trivially satisfied.

Remark 5 *We notice that, since the values $\pi_*(F|K)$ and $\pi^*(F|K)$ can assume value only in the set $\{0, 1, \Pi(E_i|H_i), i = 1, ..., n\}$, then also in this case we can build an algorithm to compute lower and upper bounds with $O(n^4)$ calls to a SAT solver. The procedure is similar to that for the unconditional case, by using the algorithm to check coherence for the conditional case [1].*

3 Some Crucial Examples

We analyze some examples of medical diagnosis (a rearrangement of those presented in [4]) to show how we can use the above theory.

Example 1. A patient feels serious generalized abdominal pains, fever and retches. The doctor puts forth the following hypotheses concerning the possible relevant disease: H_1 = *ileus*, H_2 = *peritonitis*, H_3 = *abdominal inflammation*. Moreover the doctor assumes a natural logical condition such as $H_1 \wedge H_2 = H_1 \wedge H_3 = H_2 \wedge H_3$. Correspondingly we have then five atoms $A_1 = H_1 \wedge H_2 \wedge \wedge H_3$, $A_2 = H_1 \wedge H^c_2 \wedge H^c_3$, $A_3 = H^c_1 \wedge H_2 \wedge H^c_3$, $A_4 = H^c_1 \wedge H^c_2 \wedge H_3$, $A_5 = H^c_1 \wedge H^c_2 \wedge H^c_3$.

The doctor initially gives these possibility assessments: $\Pi(H_1) = \frac{1}{2}, \Pi(H_2) = \frac{1}{3}, \Pi(H_3) = \frac{1}{5}$.

By using the algorithm discussed above we can prove that this (partial) assessment is *coherent*.

The doctor considers now the event: E = *pressing in particular points of the abdomen does not increase pain* and he gives the following relevant logical and probabilistic information $E \wedge H_1 = E \wedge H_2$, $\Pi(E|H_1) = 1$, $\Pi(E|H_1 \wedge H^c_2) = \frac{2}{5}$, $\Pi(E|H^c_1 \wedge H_2) = 0$.

Obviously, the latter assignment is coherent, since it refers to a (trivial) partition (with respect to the *conditioning* events), according to Theorem 3.

The updating of that assessment *obviously* requires that the "whole" prior and the possibilistic likelihood must be *jointly* coherent. Instead in this case coherence does not hold: it is enough to consider the relevant system associated to the restriction $\Pi(E|H_1), \Pi(H_1), \Pi(H_3)$. In fact, the following system

$$\begin{cases} \max\{x_1, x_2\} = \frac{1}{2} \\ \max\{x_1, x_4, x_5\} = \frac{1}{5} \\ x_1 = \min\{1, \max\{x_1, x_2\}\} \\ \max\{x_1, ..., x_5\} = 1 \\ x_i \geq 0\,, \end{cases}$$

admits no solution, a contradiction is obtained since from the third equation $x_1 = \frac{1}{2}$ and it contradicts the second equation.

The next example shows that it is possible to update (prior) possibility, also in unusual situations (such as when we assume that the diseases are not mutually exclusive), if coherence of the "global" (*i.e.* prior and likelihood together) assessment holds.

Example 3. A patient arrives at the hospital showing symptoms of choking. The doctor considers the following hypotheses concerning the patient situation:

$$A{=}\textit{cardiac insufficiency}$$
$$B{=}\textit{asthma attack}$$
$$C{=}\textit{respiratory problem}$$
$$D{=}\textit{cardiac insufficiency caused by asthma attack}$$

with the logical constraints $A \wedge C = \emptyset$ and $D \subset A \wedge B$. Consider the following assessment

$$\Pi(A \vee B | A \vee B \vee C) = \Pi(C | A \vee B) = 0.6,$$
$$\Pi(D|A) = \Pi(D|A \wedge B) = 0.4, \Pi(A|A \vee B) = 0.7.$$

The atoms spanned by the above events are $C_1 = A^c \wedge B^c \wedge C \wedge D^c$, $C_2 = A^c \wedge B \wedge C \wedge D^c$, $C_3 = A^c \wedge B \wedge C^c \wedge D^c$, $C_4 = A \wedge B \wedge C^c \wedge D^c$, $C_5 = A \wedge B \wedge C^c \wedge D$, $C_6 = A \wedge B^c \wedge C^c \wedge D^c$, $C_7 = A^c \wedge B^c \wedge C^c \wedge D^c$. Through the compatibility of the systems S_α we could prove that such assessment is a coherent conditional possibility.

Then, the assessment can be extended, for example, to the event $A \wedge B|A$. We need to check which are the values for the upper possibility and the lower bound, in [10] we prove that all the values inside the interval with extremes the lower and upper bounds are coherent. In Remark 4 we notice that these extremes can vary in this case in the set $\{0, 0.4, 0.6, 0.7, 1\}$.

To compute the lower bound, through the procedure implemented in the algorithm, it is easy to check that the value $\Pi(A \wedge B|A) = 0$ together with the given assessment is not coherent, while the coherence fulfils for the value $\Pi(A \wedge B|A) = 0.4$, that is the minimum coherent value for $A \wedge B|A$. For the upper possibility we can easily prove that the value 1 is coherent. Then, for $A \wedge B|A$, the coherent values are $[0.4, 1]$.

References

[1] Baioletti, M., Petturiti, D.(2009). Algorithms for possibility assessments: coherence and extension. Thecnical Report of Department of Mathematics and Informatics of Perugia University, n. 6.

[2] Bouchon-Meunier B., Coletti G., Marsala C.(2001). Conditional Possibility and Necessity. In *Technologies for Constructing Intelligent Systems*, Vol.2, 59–71, (Bouchon-Meunier, B., Gutiérrez-Rios, J., Magdalena, L., Yager, R.R.eds.) Springer, Berlin .

[3] Bouchon-Meunier B. , Coletti G., C. Marsala C. (2002). Independence and possibilistic conditioning. *Annals of Mathematics and Artificial Intelligence*, 35 107–124.

[4] Coletti G., Scozzafava R. (2000). The role of coherence in eliciting and handling imprecise probabilities and its application to medical diagnosis. *Information Science*, vol. 130 , 41–65

[5] Coletti G., Scozzafava R. (2002). *Probabilistic Logic in a Coherent Setting*. Kluwer Academic Publishers, Trends in Logic, n.15, Dordrecht - Boston - London.

[6] Coletti, G., Scozzafava, R.(2003). Conditional probability, fuzzy sets and possibility: a unifying view. *Fuzzy Sets and Systems*, vol. 144, 227-249.

[7] Coletti, G., Scozzafava, R., Vantaggi B. (2008). Possibility measures through a probabilistic inferential process *Proc. of North America Fuzzy Information Processing Society 2008* (NAFIPS 2008, New York, USA) Omnipress.

[8] Coletti, G., Scozzafava, R., Vantaggi B. (2008). Possibility measures in probabilistic inference. In: Advances in Soft computing. Tolosa (Francia), Springer.

[9] Coletti G, Vantaggi B. (2006). Possibility Theory: Conditional Independence. *Fuzzy Sets and Systems*, vol. 157 (11),1491–1513.

[10] Coletti G, Vantaggi B (2009). T-conditional possibilities: coherence and inference. *Fuzzy Sets and Systems*, 160, 306–324 .

[11] B. de Finetti, "Sul significato soggettivo della probabilità", *Fundamenta Mathematicae*, 17: 298–329, 1931 - *Engl. transl.* in: *Induction and Probability* (Eds. P. Monari, D. Cocchi), CLUEB, Bologna: 291–321, 1993.

[12] M. Delgado and S. Moral, "On the concept of possibility-probability consistency," *Fuzzy Sets and Systems* 21(3), pp. 311-318, 1987.

[13] D. Dubois, M. Grabisch, O. De Mouzon and H. Prade. A possibilistic framework for single-fault causal diagnosis under uncertainty. Int. J. General Systems, Vol. 30(2), 2001, 167-192.

[14] D. Dubois, H.T. Nguyen and H. Prade, "Possibility theory, probability and fuzzy sets: misunderstandings, bridges and gaps," In: *Fundamentals of Fuzzy Sets* (Dubois, D. Prade, H., Eds). Kluwer, Boston, Mass., The Handbooks of Fuzzy Sets Series, 343–438, 2000.

[15] D. Dubois and H. Prade, "Qualitative possibility theory and its probabilistic connections," In: *Soft Methods in Probability, Statistics and Data Analysis* (Grzegorzewski, P. et al. Eds). Physica Verlag, Heidelberg-Germany, 3–26, 2002.

[16] D. Dubois, H. Prade and P. Smets, "A definition of subjective possibility," *Operational Research and Decisions* 4, 7–22, 2003.

[17] L. Ferracuti, B. Vantaggi (2006). Independence and conditional possibilities for strictly monotone triangular norms. *International Journal of Intelligent Systems*.

[18] I. R. Goodman, H. T. Nguyen (1988). Conditional objects and the modeling of uncertainties. In: *Fuzzy Computing* (Eds. M.Gupta, T.Yamakawa), 119–138, North Holland, Amsterdam.

On General Conditional Prevision Assessments

Veronica Biazzo
Dip. Matematica e Informatica
University of Catania
vbiazzo@dmi.unict.it

Angelo Gilio
Dip. Metodi e Modelli Matematici
"Sapienza" University of Rome
gilio@dmmm.uniroma1.it

Giuseppe Sanfilippo
Dip. Scienze Statistiche e Matematiche
University of Palermo
sanfilippo@unipa.it

Abstract

In this paper we consider general conditional random quantities of the kind $X|Y$, where X and Y are finite discrete random quantities. Then, we introduce the notion of coherence for conditional prevision assessments on finite families of general conditional random quantities. Moreover, we give a compound prevision theorem and we examine the relation between the previsions of $X|Y$ and $Y|X$. Then, we give some results on random gains and, by a suitable alternative theorem, we obtain a characterization of coherence. We also propose an algorithm for the checking of coherence. Finally, we briefly examine the case of imprecise conditional prevision assessments by introducing the notions of generalized and total coherence. To illustrate our results, we consider some examples.

1 Introduction

In a recent paper ([1]) we have studied the notion of general conditional prevision $\mathbb{P}(X|Y)$, where X and Y are finite discrete random quantities. This general notion of conditional prevision has been introduced by Lad and Dickey in [5] and also discussed in [6]. In their work Lad and Dickey consider a notion of conditional prevision of the form $\mathbb{P}(X|Y)$ where both X and Y are random quantities, by generalizing the de Finetti's definition of a conditional prevision assertion $\mathbb{P}(X|H)$, where H is an event. In [5, 6] the case $\mathbb{P}(Y) = 0$ has not been considered; on the other hand, $\mathbb{P}(Y) = 0$ doesn't imply $\mathbb{P}(XY) = 0$; then $\mathbb{P}(X|Y)$ might not exist. In order to handle the case $\mathbb{P}(Y) = 0$ in [1] we have proposed a notion of coherence which integrates the definition of $\mathbb{P}(X|Y)$ given by Lad and Dickey. In particular, among other results, we have given a strong generalized compound prevision theorem. In this paper we continue our study, by considering in general conditional prevision assessments on finite families of finite discrete conditional random quantities. We introduce in general the notion of coherence; we examine the compound prevision theorem and a kind of

generalization of Bayes theorem; we obtain some results on random gains; moreover, we give some results to characterize coherence and, by exploiting them, we propose an algorithm for the checking of coherence; finally, we consider the case of imprecise conditional prevision assessments, by introducing the notions of generalized coherence and total coherence. We illustrate our results by some examples.

The paper is organized as follows: in Section 2 we give some preliminary notions and results; in Section 3 we introduce in general the notion of coherence for conditional prevision assessments; in Section 4 we generalize the compound prevision theorem and we examine the relation between $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y|X)$; in Section 5 we give some results on random gains; in Section 6 we illustrate a procedure, by proposing an algorithm, for the checking of coherence; in Section 7 we briefly examine the case of imprecise conditional prevision assessments by introducing the notions of generalized and total coherence; finally, in Section 8 we give some conclusions and comments on future work.

2 Some preliminary notions and results

We recall below two definitions given in [5, 6].

Definition 1. The conditional prevision for X given Y, denoted $\mathbb{P}(X|Y)$, is a number you specify with the understanding that you accept to engage any transaction yielding a random net gain $G = sY[X - \mathbb{P}(X|Y)]$, where s is an arbitrary real quantity.

Definition 2. Having asserted your conditional prevision $\mathbb{P}(X|Y) = \mu$, the conditional random quantity $X|Y$ is defined as

$$X|Y = XY + (1 - Y)\mu = \mu + Y(X - \mu)\,. \tag{1}$$

In [1] some critical comments and examples have been given on the previous definitions. Then, based on the notion of coherence given in [2, 4, 7, 8, 9], the following definition has been proposed

Definition 3. Given two random quantities X, Y and a conditional prevision assessment $\mathbb{P}(X|Y) = \mu$, let $G = s(X|Y - \mu) = sY(X - \mu)$ be the net random gain, where s is an arbitrary real quantity, with $s \neq 0$. Defining the event $H = (Y \neq 0)$, the assessment $\mathbb{P}(X|Y) = \mu$ is coherent if and only if: $inf\, G|H \cdot sup\, G|H \leq 0$, for every s.

Let be $X \in \mathcal{C}_X = \{x_1, \ldots, x_n\}$ and $Y \in \mathcal{C}_Y = \{y_1, \ldots, y_r\}$, with $y_k \geq 0\,,\ \forall\, k$, and $(X, Y) \in \mathcal{C} \subseteq \mathcal{C}_X \times \mathcal{C}_Y$. We denote by X^0 the subset of $\mathcal{C}_X$ such that for each $x_h \in X^0$ there exists $(x_h, y_k) \in \mathcal{C}$ with $y_k \neq 0$. Then, we set

$$x_0 = min\, X^0\,, \quad x^0 = max\, X^0\,. \tag{2}$$

Then, we have ([1])

Theorem 1. Given two finite random quantities X, Y, with $Y \geq 0$, the prevision assessment $\mathbb{P}(X|Y) = \mu$ is coherent if and only if $x_0 \leq \mu \leq x^0$.

A similar result holds for $Y \leq 0$.

3 Coherence of general conditional previsions

Given any random quantities $X_1, \ldots, X_n, Y_1, \ldots, Y_n$, based on Definitions 1 and 2 we denote by $\mathcal{M}_n = (\mu_1, \ldots, \mu_n)$ a vector of conditional previsions for "X_1 given Y_1", ..., "X_n given Y_n", where $\mu_i = \mathbb{P}(X_i|Y_i)$; then, we set $\mathcal{F}_n = \{X_1|Y_1, \ldots, X_n|Y_n\}$ and we denote by

$$\mathcal{G}_n = \sum_i^n s_i(X_i|Y_i - \mu_i) = \sum_i^n s_i Y_i(X_i - \mu_i)\,,$$

where $s_1, \ldots, s_n$ are arbitrary real quantities, the random gain associated with the pair $(\mathcal{F}_n, \mathcal{M}_n)$. We set $H_i = (Y_i \neq 0)$, $\mathcal{H}_n = H_1 \vee \cdots \vee H_n$; then, based on [2, 4, 7, 8, 9], we generalize Definition 3 by the following

Definition 4. Let $\mathbb{P}$ be a real function defined on a family $\mathcal{K}$ of conditional random quantities. $\mathbb{P}$ is said coherent if and only if, for every integer n, for every $s_1, \ldots, s_n$, and for every sub-family $\mathcal{F}_n \subseteq \mathcal{K}$, denoting by $\mathcal{M}_n = (\mu_1, \ldots, \mu_n)$ the restriction of $\mathbb{P}$ to $\mathcal{F}_n$, the following condition is satisfied

$$inf\ \mathcal{G}_n|\mathcal{H}_n \ \leq\ 0\ \leq\ sup\,\mathcal{G}_n|\mathcal{H}_n\,, \tag{3}$$

which is equivalent to $inf\ \mathcal{G}_n|\mathcal{H}_n \ \leq\ 0$, or $sup\,\mathcal{G}_n|\mathcal{H}_n \geq 0$.

We give below an example where, based on Definition 4, it is shown that in some cases do not exist finite coherent conditional prevision assessments.

Example 1. Let be given a random quantity $X \in \{-1, 1\}$, with $\mathbb{P}(X) = 0$, i.e. $P(X = -1) = P(X = 1) = \frac{1}{2}$. Of course, it is $X^2 = \mathbb{P}(X^2) = 1$; hence, the assessment $\mathbb{P}(X) = 0$ has the unique extension $\mathbb{P}(X^2) = 1$. It can be shown that the assessment $(0, 1)$ on $\{X, X^2\}$ has no finite extensions on $X|X$. In fact, let $\mathcal{M}_3 = (0, 1, \mu)$ be a prevision assessment on $\mathcal{F}_3 = \{X, X^2, X|X\}$, where $\mu = \mathbb{P}(X|X)$. By compound prevision theorem, $\mathbb{P}(XY) = \mathbb{P}(Y)\mathbb{P}(X|Y)$, it should be $\mathbb{P}(X^2) = \mathbb{P}(X)\mathbb{P}(X|X)$, that is: $1 = 0 \cdot \mu$, which has no finite solutions in the unknown μ. We will show that, for every finite quantity μ, the condition of coherence is not satisfied. In our case $H_1 = H_2 = H_3 = \Omega = \mathcal{H}_3$, so that

$$\mathcal{G}_3|\mathcal{H}_3 = \mathcal{G}_3 = s_1(X-0)+s_2(X^2-1)+s_3X(X-\mu) = (s_1-s_3\mu)X+(s_2+s_3)X^2-s_2\,;$$

then, denoting by g_1 (resp., g_2) the value of $\mathcal{G}_3$ associated with $X = -1$ (resp., $X = 1$), it is $g_1 = -s_1 + (1+\mu)s_3$, $g_2 = s_1 + (1-\mu)s_3$. Hence $s_1 < (1+\mu)s_3$ implies $g_1 > 0$, while $s_1 > (-1+\mu)s_3 = (1+\mu)s_3 - 2s_3$ implies $g_2 > 0$. Then, for every pair (s_1, s_3), with $s_3 > 0$ and $(1+\mu)s_3 < s_1 < -2s_3(1+\mu)s_3$ it is $g_1 > 0$, $g_2 > 0$; that is: $inf\ \mathcal{G}_3|\mathcal{H}_3 > 0$. Thus, the assessment $(0, 1, \mu)$ on $\{X, X^2, X|X\}$ is not coherent, for every finite μ.
We remark that, still assuming $X \in \{-1, 1\}$ and $\mathbb{P}(X) = 0$, the incoherence of the assessment $\mathbb{P}(X|X) = \mu$ can be proved by directly observing that it should be $\mathbb{P}[(X|X) - \mu] = 0$; that is $\mathbb{P}[X(X-\mu)] = \mathbb{P}(X^2 - \mu X) = 1 - \mu \cdot 0 = 0$, which is false, for every μ.

4 Compound prevision and Bayes theorems

We give below a result which generalizes the compound probability theorem to the case of n arbitrary random quantities $X_1, \ldots, X_n$.

Theorem 2. Given n random quantities $X_1, \ldots, X_n$, we have

$$\mathbb{P}(X_1 \cdots X_n) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1) \cdots \mathbb{P}(X_n|X_1 \cdots X_{n-1})\,.$$

Proof. The proof immediately follows by the compound prevision theorem; in fact, by suitably iterating the formula $\mathbb{P}(XY) = \mathbb{P}(Y)\mathbb{P}(X|Y)$, we have

$$\mathbb{P}(X_1 \cdots X_n) = \mathbb{P}(X_1 \cdots X_{n-1})\mathbb{P}(X_n|X_1 \cdots X_{n-1}) =$$

$$= \mathbb{P}(X_1 \cdots X_{n-2})\mathbb{P}(X_{n-1}|X_1 \cdots X_{n-2})\mathbb{P}(X_n|X_1 \cdots X_{n-1}) = \cdots =$$

$$= \mathbb{P}(X_1)\mathbb{P}(X_2|X_1) \cdots \mathbb{P}(X_n|X_1 \cdots X_{n-1})\,.$$

□

The following result gives a kind of generalization of Bayes theorem, by analyzing the relationship between $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y|X)$.

Theorem 3. Given two finite random quantities X, Y, with $\mathbb{P}(X) \neq 0$, we have

$$\mathbb{P}(Y|X) = \mathbb{P}(X|Y) \cdot \frac{\sum_j y_j P(Y = y_j)}{\sum_j P(Y = y_j)\mathbb{P}(X|Y = y_j)}\,.$$

Proof. We have $\mathbb{P}(XY) = \mathbb{P}(Y)\mathbb{P}(X|Y) = \mathbb{P}(X)\mathbb{P}(Y|X)$; then

$$\mathbb{P}(Y|X) = \mathbb{P}(X|Y) \cdot \frac{\mathbb{P}(Y)}{\mathbb{P}(X)} = \mathbb{P}(X|Y) \cdot \frac{\sum_j y_j P(Y = y_j)}{\sum_j P(Y = y_j)\mathbb{P}(X|Y = y_j)}\,.$$

□

Given any event E and a random quantity Y, with $\mathbb{P}(Y) \neq 0$, we have

$$\mathbb{P}(E|Y) = \frac{\mathbb{P}(Y|E)P(E)}{\mathbb{P}(Y)} = P(E) \cdot \frac{\sum_j y_j P(Y = y_j|E)}{\sum_j y_j P(Y = y_j)}\,.$$

Moreover, given two logically incompatible events A and B, we have

$$\mathbb{P}(A \vee B|Y) = \mathbb{P}(A + B|Y) = \mathbb{P}(A|Y) + \mathbb{P}(B|Y) =$$

$$= P(A) \cdot \frac{\sum_j y_j P(Y = y_j|A)}{\sum_j y_j P(Y = y_j)} + P(B) \cdot \frac{\sum_j y_j P(Y = y_j|B)}{\sum_j y_j P(Y = y_j)}\,.$$

5 Some results on random gains

In this section we deepen the notion of coherence given in Definition 4 and we obtain further theoretical results. Given any integer n, we set $J_n = \{1, \ldots, n\}$. Let be given a conditional prevision assessment $\mathcal{M}_n = (\mu_i, i \in J_n)$ on a family $\mathcal{F}_n = \{X_i|Y_i, i \in J_n\}$ of n conditional random quantities, where $\mu_i = \mathbb{P}(X_i|Y_i)$. For each subset $K \subseteq J_n$, we set $\mathcal{H}_K = \bigvee_{i \in K} H_i$; moreover, considering the sub-assessment $\mathcal{M}_K = (\mu_i, i \in K)$ on the sub-family $\mathcal{F}_K = \{X_i|Y_i, i \in K\}$, we denote by $\mathcal{G}_K$ the random gain associated with the pair $(\mathcal{F}_K, \mathcal{M}_K)$. Of course, $\mathcal{G}_n = \mathcal{G}_{J_n}$ and $\mathcal{F}_n = \mathcal{F}_{J_n}$. We denote by $\mathcal{K}$ the class of the sets $K \subseteq J_n$ which satisfy the condition $inf\, \mathcal{G}_n|\mathcal{H}_K \cdot sup\, \mathcal{G}_n|\mathcal{H}_K > 0$ for some $s_i \in \mathbb{R}, i \in J_n$. Of course, $\mathcal{K}$ may be empty. We have

Theorem 4. The class $\mathcal{K}$ is additive; that is, for every $K' \in \mathcal{K}, K'' \in \mathcal{K}$, it is $K' \cup K'' \in \mathcal{K}$. Moreover, for every $K \in \mathcal{K}$, if $K' \subset K$, then $K' \in \mathcal{K}$.

Proof. Assume that $K' \in \mathcal{K}, K'' \in \mathcal{K}$; i.e., $inf\, \mathcal{G}_n|\mathcal{H}_{K'} > 0$, $inf\, \mathcal{G}_n|\mathcal{H}_{K''} > 0$. We observe that the set of values of $\mathcal{G}_n|\mathcal{H}_{K' \cup K''}$ is the union of the set of values of $\mathcal{G}_n|\mathcal{H}_{K'}$ and $\mathcal{G}_n|\mathcal{H}_{K''}$; therefore

$$inf\, \mathcal{G}_n|\mathcal{H}_{K' \cup K''} = min\, \{inf\, \mathcal{G}_n|\mathcal{H}_{K'},\, inf\, \mathcal{G}_n|\mathcal{H}_{K''}\} > 0\,;$$

hence $K' \cup K'' \in \mathcal{K}$. Moreover, given any $K \in \mathcal{K}$ and any $K' \subset K$, as $\mathcal{H}_{K'} \subseteq \mathcal{H}_K$, the set of values of $\mathcal{G}_n|\mathcal{H}_{K'}$ is contained in the set of values of $\mathcal{G}_n|\mathcal{H}_K$ and hence $inf\, \mathcal{G}_n|\mathcal{H}_{K'} \geq inf\, \mathcal{G}_n|\mathcal{H}_K > 0$; therefore $K' \in \mathcal{K}$. □

We set

$$K_0 = \bigcup_{K \in \mathcal{K}} K\,, \quad \Gamma_0 = J_n \setminus K_0\,. \tag{4}$$

Of course, $K_0 \in \mathcal{K}$ and $\mathcal{K}$ is the power set of K_0; in conclusion, given any $K \subseteq J_n$, it is $K \setminus K_0 \neq \emptyset$, i.e. $K \notin \mathcal{K}$, if and only if $inf\, \mathcal{G}_n|\mathcal{H}_K \leq 0$. Then, we have

Theorem 5. Given a family $\mathcal{F}_n = \{X_i|Y_i, i \in J_n\}$ of n conditional random quantities and any conditional prevision $\mathcal{M}_n = (\mu_i, i \in J_n)$ on $\mathcal{F}_n$, let $(\mathcal{F}_{\Gamma_0}, \mathcal{M}_{\Gamma_0})$ be the pair associated with the subset Γ_0 defined as in (4). The conditional prevision sub-assessment $\mathcal{M}_{\Gamma_0}$ on the sub-family $\mathcal{F}_{\Gamma_0}$ is coherent.

Proof. Based on Definition 4, we have to prove that, for every $J \subseteq \Gamma_0$, with $J \neq \emptyset$, it is $inf\, \mathcal{G}_J|\mathcal{H}_J \leq 0$. Given any $J \subseteq \Gamma_0$, as $J \notin \mathcal{K}$, it is $inf\, \mathcal{G}_n|\mathcal{H}_J \leq 0$, for every $s_1, \ldots, s_n$. Moreover, $\mathcal{G}_n|\mathcal{H}_J = \mathcal{G}_J|\mathcal{H}_J + \mathcal{G}_{J_n \setminus J}|\mathcal{H}_J$; in particular, if we choose $s_i = 0$ for $i \notin J$, it is $\mathcal{G}_n|\mathcal{H}_J = \mathcal{G}_J|\mathcal{H}_J$. Then, in order the condition $inf\, \mathcal{G}_n|\mathcal{H}_J \leq 0\,, \;\; \forall\, s_1, \ldots, s_n$, be satisfied, it must be $inf\, \mathcal{G}_J|\mathcal{H}_J \leq 0$, for every s_i, $i \in J$. Therefore, the assessment $\mathcal{M}_{\Gamma_0}$ on $\mathcal{F}_{\Gamma_0}$ is coherent. □

Remark 1. We observe that $inf\, \mathcal{G}_J|\mathcal{H}_J > 0$ for some s_i, with $i \in J$, implies $inf\, \mathcal{G}_n|\mathcal{H}_J > 0$ with the same s_i, for $i \in J$, and $s_i = 0$, for $i \in J_n \setminus J$.

We give below a necessary and sufficient condition of coherence.

Theorem 6. Let be given a family $\mathcal{F}_n = \{X_i|Y_i, i \in J_n\}$ of n conditional random quantities and a conditional prevision assessment $\mathcal{M}_n = (\mu_i, i \in J_n)$ on $\mathcal{F}_n$. Moreover, let K^* be any non empty subset of J_n such that $K_0 \subseteq K^*$. The assessment $\mathcal{M}_n$ is coherent if and only if:
(i) $inf\, \mathcal{G}_n|\mathcal{H}_n \cdot sup\, \mathcal{G}_n|\mathcal{H}_n \leq 0 \;\forall\; s_i \in \mathbb{R}, i \in J_n$; (ii) $\mathcal{M}_{K^*}$ on $\mathcal{F}_{K^*}$ is coherent.

Proof. Of course, coherence of $\mathcal{M}_n$ implies (i) and (ii). Conversely, based on Definition 4, we have to prove that, for every $K \subseteq J_n$, it is $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$. We distinguish two cases: (a) $K \subseteq K^*$; (b) $K \nsubseteq K^*$. In the case (a) the condition $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$ follows from coherence of $\mathcal{M}_{K^*}$; in the case (b), $K \nsubseteq K_0$ and hence $K \notin \mathcal{K}$; therefore $inf\, \mathcal{G}_n|\mathcal{H}_K \leq 0$. Then, by reasoning as in Theorem 5, it follows $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$. Therefore $\mathcal{M}_n$ is coherent. □

We illustrate the previous result by the following

Example 2. Given a random vector (X_1, X_2, Y_1, Y_2), assume that the constituents are

$$C_1 = (X_1 = 1, X_2 = 0, Y_1 = 0, Y_2 = 1), \quad C_2 = (X_1 = 1, X_2 = 0, Y_1 = 1, Y_2 = 1),$$
$$C_3 = (X_1 = 0, X_2 = 0, Y_1 = 1, Y_2 = 1), \quad C_4 = (X_1 = 1, X_2 = 2, Y_1 = 0, Y_2 = 0),$$
$$C_5 = (X_1 = 1, X_2 = 2, Y_1 = 1, Y_2 = 0), \quad C_6 = (X_1 = 0, X_2 = 2, Y_1 = 1, Y_2 = 0).$$

Then, consider the assessment $\mathcal{M}_3 = (0, 1, 0)$ on $\mathcal{F}_3 = \{X_1|Y_1, X_2|Y_2, Y_2|X_2\}$. We observe that $\mathcal{H}_3 = (Y_1 \neq 0) \vee (Y_2 \neq 0) \vee (X_2 \neq 0) = \Omega$ and $\mathcal{G}_3|\mathcal{H}_3 = \mathcal{G}_3 = s_1 Y_1 X_1 + s_2 Y_2 (X_2 - 1) + s_3 X_2 Y_2$. The values of $\mathcal{G}_3|\mathcal{H}_3$ are

$$g_1 = -s_2\,,\ g_2 = s_1 - s_2\,,\ g_3 = -s_2\,,\ g_4 = 0\,,\ g_5 = s_1\,,\ g_6 = 0\,.$$

Now, it can be verified that $inf\, \mathcal{G}_3|H_1 \leq 0$ and $inf\, \mathcal{G}_3|H_3 \leq 0$ for all s_1, s_2, s_3, which means that $\{1, 3\} \subseteq \Gamma_0$. On the contrary, for some s_1, s_2, s_3 (e.g. for $s_2 > 0,\ s_1 < s_2$) it is $inf\, \mathcal{G}_3|H_2 \cdot sup\, \mathcal{G}_3|H_2 = -s_2(s_1 - s_2) > 0$. Thus, $\Gamma_0 = \{1, 3\}$ and $K_0 = \{2\}$. Moreover, $\mathcal{G}_{K_0}|\mathcal{H}_{K_0} = s_2 Y_2 (X_2 - \mu_2)|H_2 = -s_2$; hence the condition $inf\, \mathcal{G}_{K_0}|\mathcal{H}_{K_0} \leq 0$ is not satisfied for every s_2. This means that condition (ii) is not satisfied, i.e. the assessment $\mu_2 = 1$ on $X_2|Y_2$ is not coherent, so that $\mathcal{M}_3$ is not coherent too.
Of course, by Theorem 5, the assessment $(0, 0)$ on $\{X_1|Y_1, Y_2|X_2\}$ is coherent.

6 A procedure for checking coherence

In this section, based on a suitable alternative theorem, we characterize the coherence of conditional prevision assessments by some theoretical results; then we propose an algorithm for the checking of coherence.
Let $\mathbf{z}$, $\mathbf{s}$ and A be, respectively, a row $m-$vector, a column $n-$vector and a $m \times n-$matrix. The vector $\mathbf{z} = (z_1, \ldots, z_m)$ is said *semipositive* if $z_i \geq 0$, $\forall\, i \in J_m$ and $z_1 + \cdots + z_m > 0$. Then, we have (Gale 1960; Theorem 2.9)

Theorem 7. Exactly one of the following alternatives holds.
(i) the equality $\mathbf{z}A = 0$ has a *semipositive* solution;
(ii) the inequality $A\mathbf{s} > 0$ has a solution.

We observe that the equality $\mathbf{z}A = 0$ has a *semipositive* solution $\mathbf{z} = (z_1, \ldots, z_m)$ if and only if the equality $\mathbf{p}A = 0$ has a *semipositive* solution $\mathbf{p} = (p_1, \ldots, p_m)$ with $p_1 + \cdots + p_m = 1$.
Given two random vectors $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$, we set $(X, Y) = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$; moreover, we denote by $\mathcal{C}_{XY}$ the realm of (X, Y), that is the (finite) set of points $(\mathbf{x}, \mathbf{y}) \in \mathbf{R}^{2n}$ such that $(X = \mathbf{x}, Y = \mathbf{y}) \neq \emptyset$. We recall that $H_i = (Y_i \neq 0),\ i \in J_n,\ \mathcal{H}_n = H_1 \vee \cdots \vee H_n$; moreover, we denote by $\mathcal{C}^0_{XY} =$

$\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_m,\mathbf{y}_m)\}$, with $(\mathbf{x}_r,\mathbf{y}_r)=(x_{r1},\ldots,x_{rn},y_{r1},\ldots,y_{rn})$, $r\in J_m$, the subset of points $(\mathbf{x},\mathbf{y})$ of $\mathcal{C}_{XY}$ such that $\mathbf{y}\neq\mathbf{0}$; this means that, for any $(\mathbf{x},\mathbf{0})\in\mathcal{C}_{XY}$, it is $(\mathbf{x},\mathbf{0})\notin\mathcal{C}^0_{XY}$. Given an assessment $\mathcal{M}_n=(\mu_1,\ldots,\mu_n)$ on $F_n=\{X_1|Y_1,\ldots,X_n|Y_n\}$, we denote by C_r the constituent $(X=\mathbf{x}_r,Y=\mathbf{y}_r)$. Then, the value g_r of the random gain $\mathcal{G}_n|\mathcal{H}_n=\sum_{i=1}^n s_iY_i(X_i-\mu_i)$, associated with the constituent C_r, is given by

$$g_r=\sum_{i=1}^n s_iy_{ri}(x_{ri}-\mu_i)=\sum_{i=1}^n s_i(x_{ri}y_{ri}-\mu_iy_{ri})\,,\ r\in J_m\,.$$

We define the matrix $A=(a_{ri})$, where $a_{ri}=x_{ri}y_{ri}-\mu_iy_{ri}$, $r\in J_m$, $i\in J_n$, and the column $n-$vector $\mathbf{s}=(s_1,\ldots,s_n)^t$. If the inequality $A\mathbf{s}>0$ has a solution, this means $g_r>0$, $\forall\, r$; that is $inf\,\mathcal{G}_n|\mathcal{H}_n>0$. Then, by (the alternative) Theorem 7, the coherence condition $inf\,\mathcal{G}_n|\mathcal{H}_n\leq 0$, $\forall\, s_1,\ldots,s_n$, means that the equality $\mathbf{z}A=0$ has a *semipositive* solution $\mathbf{p}=(p_1,\ldots,p_m)$, with $\sum_{r=1}^m p_r=1$. This amounts to solvability of the following system

$$\begin{cases}\sum_{r=1}^m p_r(x_{ri}y_{ri}-\mu_iy_{ri})=0\,,\ i\in J_n,\\ \sum_{r=1}^m p_r=1\,;\ \ p_r\geq 0,\ r\in J_m.\end{cases}\tag{5}$$

Remark 2. Given any $K\subset J_n$, we denote by $A_K=(a_{ri})$ the sub-matrix of A such that $i\in J_n$ and r such that $C_r\subseteq\mathcal{H}_K$. By the same alternative theorem, we have that the condition $inf\,\mathcal{G}_n|\mathcal{H}_K\leq 0$, $\forall\, s_1,\ldots,s_n$, means that the inequality $A_K\mathbf{s}>0$ has no solutions, or equivalently that the equality $\mathbf{p}_KA_K=0$ has a semipositive solution $\mathbf{p}_K=(p_r,r:C_r\subseteq\mathcal{H}_K)$; i.e., the following system is solvable

$$\begin{cases}\sum_{r:C_r\subseteq\mathcal{H}_K} p_r(x_{ri}y_{ri}-\mu_iy_{ri})=0\,,\ i\in J_n,\\ \sum_{r:C_r\subseteq\mathcal{H}_K} p_r=1\,;\ \ p_r\geq 0,\ r:C_r\subseteq\mathcal{H}_K.\end{cases}\tag{6}$$

We observe that, denoting by $(x_j^{(i)},y_j^{(i)})$ the generic possible value of (X_i,Y_i), the system (5) can be equivalently rewritten as

$$\begin{cases}\sum_j x_j^{(i)}y_j^{(i)}\sum_{r:C_r\subseteq(X_i=x_j^{(i)},Y_i=y_j^{(i)})}p_r=\mu_i\sum_j y_j^{(i)}\sum_{r:C_r\subseteq(Y_i=y_j^{(i)})}p_r,\ \ i\in J_n,\\ \sum_{r=1}^m p_r=1\,;\ \ p_r\geq 0,\ r\in J_m.\end{cases}\tag{7}$$

Notice that, in probabilistic terms, we have the following interpretations

$$\begin{aligned}&p_r=P(C_r|\mathcal{H}_n)=P[(X=\mathbf{x}_r,Y=\mathbf{y}_r)|\mathcal{H}_n]\,;\\ &\textstyle\sum_{r:C_r\subseteq(X_i=x_j^{(i)},Y_i=y_j^{(i)})}p_r=P[(X_i=x_j^{(i)},Y_i=y_j^{(i)})|\mathcal{H}_n]\,;\\ &\textstyle\sum_{r:C_r\subseteq(Y_i=y_j^{(i)})}p_r=P[(Y_i=y_j^{(i)})|\mathcal{H}_n]\,;\end{aligned}\tag{8}$$

hence, system (7) can be looked at

$$\Pr(X_iY_i|\mathcal{H}_n)=\mu_iPr(Y_i|\mathcal{H}_n),\ i\in J_n\,;\ \ P(\mathcal{H}_n|\mathcal{H}_n)=1\,.\tag{9}$$

Now, assuming that system (7) is solvable, we denote by S its (non empty) set of solutions. Given any $\mathbf{p}=(p_1,\ldots,p_m)\in S$, we set

$$\Phi_j(\mathbf{p})=\sum_{r:C_r\subseteq H_j}p_r,\ M_j=max_{\mathbf{p}\in S}\,\Phi_j(\mathbf{p}),\ j\in J_n;\ \ I_0=\{j\in J_n:M_j=0\}.\tag{10}$$

Of course, solvability of system (7) implies $I_0 \subset J_n$. Given any $K \subseteq J_n$, we denote by $(\mathcal{F}_K, \mathcal{M}_K)$ the pair associated with K and by $\mathcal{G}_K|\mathcal{H}_K$ (resp., by $(\mathcal{S}_K)$) the random gain (resp., the system) associated with $(\mathcal{F}_K, \mathcal{M}_K)$.
Of course, $\mathcal{G}_n = \mathcal{G}_{J_n}$ and $\mathcal{F}_n = \mathcal{F}_{J_n}$. We have

Theorem 8. Assume that system (7) is solvable; moreover, let I_0 be defined as in (10). Then, given any $K \subset J_n$ such that $K \setminus I_0 \neq \emptyset$, the system $(\mathcal{S}_K)$ is solvable; that is $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$. Moreover, the sub-assessment $\mathcal{M}_{J_n \setminus I_0}$ on the sub-family $\mathcal{F}_{J_n \setminus I_0}$ is coherent.

Proof. Given any $j \in K \setminus I_0$ there exists a solution $\mathbf{p}^{(j)} = (p_1^{(j)}, \ldots, p_m^{(j)}) \in S$ such that $\Phi_j(\mathbf{p}^{(j)}) > 0$; moreover

$$\sum_{r:C_r \subseteq \mathcal{H}_K} p_r^{(j)} \geq \sum_{r:C_r \subseteq H_j} p_r^{(j)} = \Phi_j(\mathbf{p}^{(j)}) > 0\,.$$

Hence, $\mathbf{p}^{(j)}$ is a solution of the following system related with system (7)

$$\begin{cases} \sum_j x_j^{(i)} y_j^{(i)} \sum_{r:C_r \subseteq (X_i = x_j^{(i)}, Y_i = y_j^{(i)})} p_r = \mu_i \sum_j y_j^{(i)} \sum_{r:C_r \subseteq (Y_i = y_j^{(i)})} p_r, \quad i \in K, \\ \sum_{r:C_r \subseteq \mathcal{H}_K} p_r > 0\,; \;\; p_r \geq 0,\; r \in J_m. \end{cases} \tag{11}$$

As it can be verified, the solvability of the system (11) is equivalent to solvability of the system $(\mathcal{S}_K)$; that is, by the alternative theorem, to satisfiability of the condition $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$. In particular, the condition $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$ holds for every $K \subseteq J_n \setminus I_0$ and this amounts to coherence of $\mathcal{M}_{J_n \setminus I_0}$. □

By the previous result, we obtain

Theorem 9. Let be given a family $\mathcal{F}_n = \{X_i|Y_i, i \in J_n\}$ of n conditional random quantities and a conditional prevision assessment $\mathcal{M}_n = (\mu_i, i \in J_n)$ on $\mathcal{F}_n$. Moreover, let K^* be any non empty subset of J_n such that $I_0 \subseteq K^*$. The assessment $\mathcal{M}_n$ is coherent if and only if:
(i) the system (7) is solvable; (ii) $\mathcal{M}_{K^*}$ on $\mathcal{F}_{K^*}$ is coherent.

Proof. Of course, coherence of $\mathcal{M}_n$ implies conditions (i) and (ii). Conversely, based on Definition 4, we have to prove that $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0, \forall\, K \subseteq J_n$. We observe that, by (i), it is $inf\, \mathcal{G}_n|\mathcal{H}_n \leq 0$ and $I_0 \subset J_n$. We distinguish two cases: (a) $K \subseteq K^*$; (b) $K \nsubseteq K^*$. In the case (a) the condition $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$ follows from coherence of $\mathcal{M}_{K^*}$; in the case (b) the condition $inf\, \mathcal{G}_K|\mathcal{H}_K \leq 0$ follows by Theorem 8, as $K \setminus K^* \neq \emptyset$. □

Remark 3. We recall that, for each $r \in J_m$, C_r represents the constituent $(X = \mathbf{x}_r, Y = \mathbf{y}_r)$; hence, given any $K \subseteq J_n$, with $K \setminus I_0 \neq \emptyset$, for each r such that $C_r \subseteq (Y_i = y_j^{(i)})$, $i \in K$, we have $C_r \subseteq \mathcal{H}_K$. Hence, in system (11) for all the variables p_r's it is $C_r \subseteq \mathcal{H}_K$ and the condition $r \in J_m$ can be replaced by $r : C_r \subseteq \mathcal{H}_K$. It follows that, by defining

$$\lambda_r = \frac{p_r}{\sum_{r:C_r \subseteq \mathcal{H}_K} p_r}\,, \;\; \forall\, r : C_r \subseteq \mathcal{H}_K\,,$$

the system (11) can be rewritten as the following one

$$\begin{cases} \sum_j x_j^{(i)} y_j^{(i)} \sum_{r:C_r \subseteq (X_i = x_j^{(i)}, Y_i = y_j^{(i)})} \lambda_r = \mu_i \sum_j y_j^{(i)} \sum_{r:C_r \subseteq (Y_i = y_j^{(i)})} \lambda_r, \quad i \in K, \\ \sum_{r:C_r \subseteq \mathcal{H}_K} \lambda_r = 1\,; \;\; \lambda_r \geq 0,\; r : C_r \subseteq \mathcal{H}_K. \end{cases} \tag{12}$$

Moreover, concerning system (6), as $s_1, \ldots, s_n$ are arbitrary, by choosing $s_i = 0$, $\forall\, i \in J_n \setminus K$, the system obtained by (6), with $i \in J_n$ replaced by $i \in K$, is equivalent to system (11). As a consequence: (i) $K_0 = I_0$; (ii) Theorems 5 and 8 are equivalent; (iii) Theorems 6 and 9 are equivalent too.

We observe that, if $K \subseteq I_0$, nothing can be said about the solvability of system $(\mathcal{S}_K)$, which requires a direct checking, by starting with $K = I_0$.
Based on Theorems 8 and 9, we can use the algorithm below for the checking of coherence.

Algorithm 1. Let be given a conditional prevision assessment $\mathcal{M}_n = (\mu_1, \ldots, \mu_n)$ on $\mathcal{F}_n = \{X_1|Y_1, \ldots, X_n|Y_n\}$.
Step 1. Check the solvability of system (7); if the system is not solvable, then $\mathcal{M}_n$ is not coherent.
Step 2. If the system is solvable, determine I_0; if $I_0 = \emptyset$, then $\mathcal{M}_n$ is coherent.
Step 3. If $I_0 \neq \emptyset$, then determine the pair $(\mathcal{F}_{I_0}, \mathcal{M}_{I_0})$; replace the pair $(\mathcal{F}_n, \mathcal{M}_n)$ by $(\mathcal{F}_{I_0}, \mathcal{M}_{I_0})$ and repeat the previous steps.

As we can see, using the algorithm above, we can check coherence of the assessment $\mathcal{M}_n$ on $\mathcal{F}_n$ in a finite number of iterations. If the initial system is solvable, a suitable sequence of sets $I_0^{(1)}, \ldots, I_0^{(t)}$ is computed. We have two cases: (a) if $\mathcal{M}_n$ is coherent, it is $t \leq n$ and $I_0^{(t)} = \emptyset$; (b) if $\mathcal{M}_n$ is not coherent, it is $t \leq n-1$ and $I_0^{(t)} \neq \emptyset$. We give an example to illustrate Algorithm 1.

Example 3. *(we continue Example 2)* Concerning the assessment $\mathcal{M}_3 = (0, 1, 0)$ on the family $\mathcal{F}_3 = \{X_1|Y_1, X_2|Y_2, Y_2|X_2\}$, with each constituent C_r, we associate a variable p_r, $r = 1 \ldots, 6$. Then, based on Algorithm 1, we check the solvability of the initial system given below.

$$\begin{cases} 0(p_1+p_3+p_4+p_6)+1(p_2+p_5) = 0(0(p_1+p_4)+1(p_2+p_3+p_5+p_6)), \\ 0 = 1(0(p_4+p_5+p_6)+1(p_1+p_2+p_3)), \\ 0 = 0(0(p_1+p_2+p_3)+2(p_4+p_5+p_6)), \\ \sum_{r=1}^{6} p_r = 1, \;\; p_r \geq 0, \; r = 1, \ldots, 6, \end{cases} \tag{13}$$

which can be written

$$\begin{cases} p_2 + p_5 = 0\,, \;\; p_1 + p_2 + p_3 = 0\,, \;\; 0 = 0\,, \\ \sum_{r=1}^{6} p_r = 1, \;\; p_r \geq 0, \; r = 1, \ldots, 6. \end{cases} \tag{14}$$

Each vector $\mathbf{p} = (p_1, \ldots, p_6)$, with $p_1 = p_2 = p_3 = p_5 = 0, p_4 + p_6 = 1$, is a solution of this system. We have

$$\Phi_1(\mathbf{p}) = p_2 + p_3 + p_5 + p_6\,, \;\; \Phi_2(\mathbf{p}) = p_1 + p_2 + p_3\,, \;\; \Phi_3(\mathbf{p}) = p_4 + p_5 + p_6\,,$$

hence $M_1 > 0$, $M_2 = 0$, $M_3 > 0$. Then, $I_0 = \{2\}$ and we have to check the coherence of the assessment $\mu_2 = \mathbb{P}(X_2|Y_2) = 1$. As conditionally on $(Y_2 \neq 0)$ the unique possible value of X_2 is 0, it must be $\mathbb{P}(X_2|Y_2) = 0$; hence, by the algorithm it results that the assessment $\mathcal{M}_3$ is not coherent. Of course, by Theorem 8, the sub-assessment $(0, 0)$ on $\{X_1|Y_1, Y_2|X_2\}$ is coherent.

7 Imprecise conditional prevision assessments

In this section we briefly examine imprecise conditional prevision assessments; we introduce below the notions of generalized coherence and total coherence.

Definition 5. Let be given any random quantities $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ and a set $\mathcal{S} \subseteq \mathbf{R}^n$. With each point $\mathcal{M}_n = (\mu_1, \ldots, \mu_n) \in \mathcal{S}$ we associate the family $\mathcal{F}_n = \{X_1|Y_1, \ldots, X_n|Y_n\}$, where $X_i|Y_i = \mu_i + Y_i(X_i - \mu_i)$, $i \in J_n$. We say that the set $\mathcal{S}$ is coherent in a generalized sense (*g-coherent*) if and only if there exists $\mathcal{M}_n \in \mathcal{S}$ which is a coherent conditional prevision assessment on $\mathcal{F}_n$.
We say that the set $\mathcal{S}$ is *totally coherent* if and only if, for every $\mathcal{M}_n \in \mathcal{S}$, $\mathcal{M}_n$ is a coherent conditional prevision assessment on $\mathcal{F}_n$.

Of course, total coherence implies g-coherence.

Given a family of n conditional random quantities $\mathcal{F}_n = \{X_1|Y_1, \ldots, X_n|Y_n\}$, we assume $(X_i, Y_i) \in \mathcal{C}_i\,, \quad i \in J_n$; moreover, for each i, we set $\mathcal{C}_i^0 = (x, y) \in \mathcal{C}_i : y \neq 0$ and $H_i = (Y_i \neq 0)$. For each i, we denote by X_i^0 the set of values of X_i such that for each $x \in X_i^0$ there exists a possible value y of Y_i such that $(x, y) \in \mathcal{C}_i^0$. Moreover, we set $m_i = min\, X_i^0\,, \quad M_i = max\, X_i^0\,, \quad i \in J_n$. We recall that, assuming $Y_i \geq 0$, or $Y_i \leq 0$, the assessment $\mathbb{P}(X_i|Y_i) = \mu_i$ is coherent if and only if $m_i \leq \mu_i \leq M_i$. We set $I = [m_1, M_1] \times \cdots \times [m_n, M_n]$. Then, we have the following result which concerns the total coherence of I.

Theorem 10. Let be given a conditional prevision assessment $\mathcal{M}_n = (\mu_1, \ldots, \mu_n)$ on a family $\mathcal{F}_n = \{X_1|Y_1, \ldots, X_n|Y_n\}$, where for each i it is $Y_i \geq 0$, or $Y_i \leq 0$. Moreover, assume that $H_iH_j = \emptyset$, for each $i \neq j$. Then, the assessment $\mathcal{M}_n$ is coherent if and only if $m_i \leq \mu_i \leq M_i$ for every i; that is, I is totally coherent.

Proof. We set $G_i = s_iY_i(X_i - \mu_i),\ i \in J_n$; then

$$\mathcal{G}_n = G_1 + \cdots + G_n = H_1G_1 + \cdots + H_nG_n\,,$$

where $s_1, \ldots, s_n$ are arbitrary real numbers. Of course, for each i, the condition $inf\ G_i|H_i \leq 0\ \forall\, s_i$ is satisfied if and only if $m_i \leq \mu_i \leq M_i$. Then, recalling that $\mathcal{H}_n = H_1 \vee \cdots \vee H_n$, from the hypothesis $H_iH_j = \emptyset$ for $i \neq j$, it follows

$$\mathcal{G}_n|\mathcal{H}_n = \begin{cases} G_1|H_1, & H_1\ true, \\ \ldots\ldots & \ldots\ldots \\ G_n|H_n, & H_n\ true. \end{cases}$$

Then

$$inf\ \mathcal{G}_n|\mathcal{H}_n = min\ \{inf\ G_i|H_i\,,\ i \in J_n\}\,,$$

and the condition $inf\ \mathcal{G}_n|\mathcal{H}_n \leq 0\ \forall\, s_1, \ldots, s_n$, is satisfied if and only if it is satisfied the condition $inf\ G_i|H_i \leq 0\ \forall\, s_i$, $i \in J_n$; that is $m_i \leq \mu_i \leq M_i\ \forall\, s_i$, $i \in J_n$. Of course, a similar reasoning can be applied to each sub-family of $\mathcal{F}_n$; hence I is totally coherent. □

We illustrate the previous result by the following

Example 4. Assume that the random vector $(X_1, X_2, X_3, Y_1, Y_2, Y_3)$ has the following possible values

$$(1, 1, 1, 1, 0, 0)\,, \quad (-1, -1, -1, 1, 0, 0)\,, \quad (1, 1, 1, 0, 1, 0)\,,$$

$$(-1,-1,-1,0,1,0)\,,\quad (1,1,1,0,0,1)\,,\quad (-1,-1,-1,0,0,1)\,;$$

moreover, let $\mathcal{M} = (\mu_1, \mu_2, \mu_3)$ a conditional prevision assessment on $\mathcal{F}_3 = \{X_1|Y_1, X_2|Y_2, X_3|Y_3\}$. We observe that $[m_i, M_i] = [-1, 1]$, $i = 1, 2, 3$, and $I = [-1, 1]^3$. Moreover, we have the following values for the random gain $\mathcal{G}_3|\mathcal{H}_3$

$$s_1(1-\mu_1)\,,\ -s_1\mu_1\,,\ s_2(1-\mu_2)\,,\ -s_2\mu_2\,,\ \ s_3(1-\mu_3)\,,\ -s_3\mu_3\,.$$

As it can be easily verified, the condition $min\,\mathcal{G}_3|\mathcal{H}_3 \leq 0\,,\ \forall\, s_1, s_2, s_3$, is satisfied if and only if $-1 \leq \mu_i \leq 1$, $i = 1, 2, 3$; of course, a similar reasoning can be applied to each subfamily of $\mathcal{F}_3$. Hence the interval $I = [-1, 1]^3$ is totally coherent.

8 Conclusions

In this paper we have introduced the notion of coherence for conditional prevision assessments on finite families of general conditional random quantities. Moreover, we have examined the compound prevision theorem and the relation between $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y|X)$. Then, we have given some theoretical results on random gains and, based on a suitable alternative theorem, we have given a characterization of coherence. We have also proposed an algorithm for the checking of coherence. Finally, we have introduced the notions of generalized and total coherence; then, we have briefly examined the case of imprecise conditional prevision assessments. To illustrate our results we have considered some examples. Future work should concern the deepening of the case of imprecise prevision assessments.

References

[1] Biazzo V., Gilio A., and Sanfilippo G., On general conditional random quantities, *Proc. of Sixth Intern. Symposium on Impr. Probability: Theories and Applications (ISIPTA'09*, Durham University, United Kingdom, July 14 - 18, 2009.

[2] de Finetti B., Teoria delle probabilità, 2 voll., Ed. Einaudi, Torino, 1970 (translation in english: Theory of Probability 1 (2), Chichester, Wiley, 1974 (1975)).

[3] Gale, D. 1960. The theory of linear economic models, McGraw-Hill, New York.

[4] Holzer S., On coherence and conditional prevision, *Boll. Un. Mat. Ital.* 4 (6), 441-460, 1985.

[5] Lad F., and Dickey J. M., A general theory of conditional prevision, $P(X|Y)$, and the problem of state-dependent preferences, *Economic Decision-Making: Games, Econometrics and Optimization, Essays in Honor of Jaques Dreze*, J.J. Gabsewicz, J.F. Richard, and L.A. Wolsey (eds.), Amsterdam: North Holland, 369-383, 1990.

[6] Lad F., *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. New York, Wiley, 1996.

[7] Lehman R. S., On confirmation and rational betting, *The Journal of Symbolic Logic* 20: 251-262, 1955.

[8] Regazzini E., Finitely additive conditional probabilities, *Rend. Sem. Mat. Fis. Milano* 55, 69-89, 1985.

[9] Williams P.M., Notes on conditional previsions, Technical report, University of Sussex, 1975. Reprinted in a revised form in: *International Journal of Approximate Reasoning*, 44(3):366-383, 2007.

Merging Different Probabilistic Information Sources through a New Discrepancy Measure

Andrea Capotorti, Giuliana Regoli, Francesca Vattari
Dipartimento di Matematica e Informatica
Università di Perugia - Italy
{capot,regoli,francesca.vattari}@dipmat.unipg.it

Abstract

In this paper we are going to show a new way to find a consistent compromise among different opinions expressed through conditional probability assessments. This procedure will profit from the nice properties of a discrepancy measure among partial conditional probability assessments already introduced to adjust incoherent assessments. The use of such discrepancy for this further goal will be described through exemplifying examples.
Keywords: Belief revision and inconsistency handling, information merging, coherent conditional probability assessments, inference.

1 Introduction

The need of finding a consistent compromised among different opinions is an actual problem in all the ambit where information is mainly expressed through expertise or when there is the need to join different sources. Of course there are several possible way to express the available information and consequentially to operate. For example, limitedly to the field of probabilistic approaches, aggregation is deeply studied both in precise (see e.g. [10, 12, 21, 23]) and imprecise (see e.g. [11, 19, 18, 22]) frameworks. Some of the proposed aggregation rules are solely based on the assessed values, others rely on auxiliary over structures, e.g. second order assessments or risk neutral probabilities. Our choice is in between: once a specific "distance" among probability distributions is chosen, then the aggregation proceeds "alone" by working only on the assessed values.

In this paper we focus on the specific field of knowledge expressed through partial conditional probability assessments. In this area we recently introduced a procedure to correct inconsistent evaluations, see [4]. This procedure is based on a discrepancy measure among partial conditional probabilities derived by a particular scoring rule. Such a scoring rule is inspired by the one introduced by Lad in [17] for unconditional probability distributions, and adapted to partial conditional frameworks. This permits a behavioral interpretation and a probabilistic justification of the correction procedure, differently from other similar proposal, e.g. [16], that mainly rely on purely geometrical interpretations. By

profiting of the same discrepancy measure, we can now propose a way to merge different opinions. In fact the solution will be the "closest" coherent evaluation to the given disparate values.

We face two different kind of merging: in the first we tackle with evaluations given on overlapping domains, i.e. different probability values can be given on the same conditional events; in the second we join together evaluations given on separate domains, i.e. each source of information is given on a specific set of conditional events.

2 Basic notions

The different sources of information, that could represent expert's opinions and/or knowledge bases, will be indexed by a subscript index s varying on a finite set S.

We formalize the domain of the different evaluations through finite families of conditional events of the type $\mathcal{E}_s = [E1_s|H1_s, \ldots, En_s|Hn_s]$, $s \in S$. The events Ei_s's usually represent the situations under consideration in the source s, while the Hi_s's usually represent the different contexts, or scenarios, under which the Ei_s's are evaluated.

The basic events $E1_s, \ldots, En_s, H1_s, \ldots, Hn_s$ can be endowed with logical constraints, that represent dependencies among particular configurations of them (e.g. incompatibilities, implications, partial or total coincidences, etc.).

In the following Ei_sHi_s will denote the logical connection "Ei_s and Hi_s", $\neg Ei_s$ will indicate "not Ei_s" and the event $H_s^0 = \bigvee_{i=1}^{n} Hi_s$ will represent the whole set of contexts taken under consideration in the source of information $s \in S$.

By the basic events $E1_s, \ldots, En_s, H1_s, \ldots, Hn_s$, it is possible to span a sample space $\Omega_s = \{\omega_{1_s}, \ldots, \omega_{k_s}\}$, where ω_{j_s} represents generic atoms, in some context named "possible worlds ". Note that the sample space Ω_s, together with H_s^0, are not part of the assessment but only auxiliary tools.

The numerical part of the different assessments can be elicited either through precise numerical values $\mathbf{p}_s = (p1_s, \ldots, pn_s)$ thought as honest evaluation of the probabilities $P(Ei_s|Hi_s)$, $i = 1, \ldots, n_s$, or through interval values $\mathbf{p}_s = ([lb1_s, ub1_s], \ldots, [lbn_s, ubn_s])$ thought as honest ranges for the probabilities $P(Ei_s|Hi_s)$, $i = 1_s, \ldots, n_s$.

Any single assessment $(\mathcal{E}_s, \mathbf{p}_s)$, $s \in S$, is supposed to be consistent, i.e. coherent. For precise assessments, coherence requires the existence of at least a probability distribution that induces the assessed pi_s. For interval assessments, coherence requires the existence of a class of probability distributions that induce probability values for the $Ei_s|Hi_s$ inside the ranges $[lbi_s, ubi_s]$ and, at the same time, each lower (lbi_s) or upper (ubi_s) bound is actually reached through one of such distributions. For a complete and rigorous description of such notions the reader can refer to the exhaustive treatise [9].

When the different evaluations are merged, we get a unique assessment with repetitions, i.e. conditional events with different absolute frequencies, and the numerical part with both precise and imprecise values. To distinguish the whole merged assessments by its components we simply ignore the indexes $s \in S$, so that we deal with the domain $\mathcal{E} = [E1|H1, \ldots, En|Hn] = \bigcup_{s \in S} \mathcal{E}_s$ with associated assessment $\mathbf{p} = ([lp1, up1], \ldots, [lpn, upn]) = \bigcup_{s \in S} \mathbf{p}_s$.

The whole set of basic events $E1, \ldots, En, H1, \ldots, Hn$ span a unique sample space $\Omega = \{\omega_1, \ldots, \omega_k\}$, that actually turns out to be a refinement of each Ω_s.

Usually such merged assessment $(\mathcal{E}, \mathbf{p})$ is inconsistent, i.e. there is not a set of probability distributions over the sample space Ω fulfilling all the constraints induced by the different assessments. For this we want to find a consistent conditional probability assessment over $\mathcal{E}$ that will represent a compromise among the $(\mathcal{E}_s, \mathbf{p}_s)$.

The fact that the whole assessment $(\mathcal{E}, \mathbf{p})$ is the result of a merging action does not matter, we can treat it as a generic incoherent partial conditional, precise and/or imprecise, probability assessment. In fact, the possible multiplicity of some conditional event $Ei|Hi$ in $\mathcal{E}$ can be simply treated as peculiar logical relations. Hence, the searched compromise solution can be simply detected as the closest coherent assessment to $(\mathcal{E}, \mathbf{p})$.

Closeness notion implies the choice of some kind of distance, and we will profit from the aforementioned discrepancy measure. Before to introduce it again, for the sake of comprehensiveness, we need some further auxiliary notion.

Every probability distribution $\alpha : \mathcal{P}(\Omega) \to \mathbb{R}$ corresponds to a non-negative vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]$, with $\alpha_j = \alpha(\omega_j)$, so that for every event $E \in \mathcal{P}(\Omega)$ it results $\alpha(E) = \sum_{\omega_j \subseteq E} \alpha_j$.
We need to introduce a nested hierarchy among probability distribution sets:

- let $\mathcal{A} = \left\{ \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k], \sum_1^k \alpha_i = 1, \alpha_j \geq 0, j = 1, \ldots, k \right\}$ represents the whole set of probability distributions on Ω;
- let $\mathcal{A}_0 = \left\{ \boldsymbol{\alpha} \in \mathcal{A} | \alpha(H^0) = 1 \right\}$ be the subset of probability distributions on Ω that concentrate all the probability mass on the contemplated scenarios;
- let $\mathcal{A}_1 = \{ \boldsymbol{\alpha} \in \mathcal{A}_0 | \alpha(H_i) > 0, i = 1, \ldots, n \}$ be the subset of probability distributions on Ω that give positive probability to every scenario;
- let $\mathcal{A}_2 = \{ \boldsymbol{\alpha} \in \mathcal{A}_1 | 0 < \alpha(E_i H_i) < \alpha(H_i), i = 1, \ldots, n \}$ be the subset of probability distributions that avoid boundary values $\{0, 1\}$ for the conditional probabilities.

It is easy to see that the sets $\mathcal{A}_i$ are convex sets and $\mathcal{A}_0$ is the closure of $\mathcal{A}_2$ (and $\mathcal{A}_1$) in the usual topology.
Note that in conditional frameworks the focusing on $\mathcal{A}_0$ is commonly done to avoid unpleasant consequences. See Walley[24] about Avoiding Uniform Loss assessments or Holzer[14] about the Principle of Conditional Coherence.

Every probability distribution $\boldsymbol{\alpha} \in \mathcal{A}_1$ generates a coherent assessment $\mathbf{q}_{\boldsymbol{\alpha}}$ on $\mathcal{E}$ through the usual formula

$$q_{\alpha_i} = \frac{\sum\limits_{\omega_j \subseteq E_i H_i} \alpha_j}{\sum\limits_{\omega_j \subseteq H_i} \alpha_j} \qquad \forall i = 1, \ldots, n \tag{1}$$

Note that $\mathbf{q}_{\boldsymbol{\alpha}}$ is a continuous function of $\boldsymbol{\alpha}$ when $\boldsymbol{\alpha} \in \mathcal{A}_1$.
When $\boldsymbol{\alpha} \in \mathcal{A}_0$, previous formula (1) defines $\mathbf{q}_{\boldsymbol{\alpha}}$ only on

$$\mathcal{E}_{\boldsymbol{\alpha}} := \{ E_i | H_i \in \mathcal{E}, \alpha(H_i) > 0 \}. \tag{2}$$

Coherence of $\mathbf{q}_{\boldsymbol{\alpha}}$ is guaranteed by the theorem of Coletti (1994)[7].

Associated to any (coherent or not) assessment $\mathbf{p} \in (0,1)^n$ over $\mathcal{E} = [E1|H1, \ldots, En|Hn]$ we can introduce a scoring rule

$$S(\mathbf{p}) := \sum_{i=1}^{n} |EiHi| \ln p_i + \sum_{i=1}^{n} |\neg EiHi| \ln(1 - p_i) \tag{3}$$

with $|\cdot|$ indicator function of unconditional events.

Note that such scoring rule is not defined for boundary values 0 or 1 for the assessed probabilities. This is of course a limitation in our approach but we anyhow believe in its significance. In particular, if any of the pi_s, or of the lbi_s, or of the ubi_s, is 0 or 1, they can be maintained fixed in their values, if this of course will not induce any evident contradiction. This can be legitimated by the fact that, if the source had a so strong belief to asses such extreme values, it is reasonable to suppose it does not want to reach a compromise for them.

Such score $S(\mathbf{p})$ is an "adaptation" of the *total-log* "proper scoring rule" for probability distributions proposed by Lad in [17](pag. 355). We have extended it to partial and conditional probability assessments.

The motivation of such a score derives from the fact that a conditional event $Ei|Hi$ is a three-valued logical entity, partitioning Ω in three parts: the atoms satisfying $EiHi$ and thus verifying the conditional, those satisfying $\neg EiHi$, thus falsifying the conditional, and those not fulfilling the context Hi, to which the conditional may not be applied at all. Hence the assessor of $\mathbf{p}$ "loses less" the higher are the probabilities assessed for events that are verified, and at the same time, the lower are the probabilities assessed for those that are not verified. The values assessed on events that turn out to be undetermined do not influence the score. In fact the realization of the random value $S(\mathbf{p})$ when the atom ω_j occurs is

$$S_j(\mathbf{p}) = \sum_{EiHi \supseteq \omega_j} \ln p_i + \sum_{\neg EiHi \supseteq \omega_j} \ln(1 - p_i). \tag{4}$$

The choice of a scoring rule closely related to the usual logarithmic one, apart from its useful mathematical properties, is motivated by its strict connection with the well known principle of *minimum cross-entropy* (see e.g. [20]). Hence we follow the paradigm of *informational economy*, and our approach is strictly related to that of [15] but with different assumptions and techniques.

Note moreover that the concurrent involvement in the score (3) of the events that turn out to be true and those that turn out to be false, modifies the peculiar property of the usual logarithmic scoring rule to depend only on the true ones.

We have all the elements now to introduce the "discrepancy" between an assessment $\mathbf{p}$ over $\mathcal{E}$ and a distribution $\boldsymbol{\alpha} \in \mathcal{A}_2$, with respect to its induced conditional coherent assessment $\mathbf{q}_{\boldsymbol{\alpha}}$, as

$$\Delta(\mathbf{p}, \boldsymbol{\alpha}) := E_{\boldsymbol{\alpha}}(S(\mathbf{q}_{\boldsymbol{\alpha}}) - S(\mathbf{p})) = \sum_{j=1}^{k} \alpha_j [S_j(\mathbf{q}_{\boldsymbol{\alpha}}) - S_j(\mathbf{p})]\,. \tag{5}$$

The need of a "discrepancy" instead of a usual "distance" (or better "divergence") is motivated by the (general) non-convexity of the coherent set of conditional assessment (see [13] and the next Ex.1 we borrowed from it). It is easy

to see that

$$\Delta(\mathbf{p},\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha(E_iH_i)\ln(\frac{q_i}{p_i}) + \alpha(E_i^cH_i)\ln(\frac{1-q_i}{1-p_i}) \quad (6)$$

$$= \sum_{i=1}^{n} \alpha(H_i)\left(q_i\ln(\frac{q_i}{p_i}) + (1-q_i)\ln(\frac{1-q_i}{1-p_i})\right) \quad (7)$$

The restriction to the distributions $\boldsymbol{\alpha}$ in $\mathcal{A}_2$ is because only there the scoring rule $S(\mathbf{q}_{\boldsymbol{\alpha}})$ is properly defined. Anyhow, it is possible to extend by continuity the previous definition of $\Delta(\mathbf{p},\boldsymbol{\alpha})$ to any distribution $\boldsymbol{\alpha}$ in $\mathcal{A}_0$ trough the expression

$$\Delta(\mathbf{p},\boldsymbol{\alpha}) := \sum_{i|\alpha(H_i)>0} \alpha(H_i)\left(q_i\ln(\frac{q_i}{p_i}) + (1-q_i)\ln(\frac{1-q_i}{1-p_i})\right) \quad (8)$$

adopting the usual convention $0\ln(0) = 0$.

Such discrepancy $\Delta(\mathbf{p},\boldsymbol{\alpha})$ behaves analogously to other usual Bregman divergences[1] (see [2]). In fact in [5] we formally prove that the following properties hold.

- $\Delta(\mathbf{p},\boldsymbol{\alpha}) \geq 0 \quad \forall \boldsymbol{\alpha} \in \mathcal{A}$;
- $\Delta(\mathbf{p},\boldsymbol{\alpha}) = 0$ iff $\mathbf{p}_{|\mathcal{E}_\alpha} \equiv \mathbf{q}_{\boldsymbol{\alpha}}$;
- $\Delta(\mathbf{p},\cdot)$ is convex on $\mathcal{A}_2$;
- $\Delta(\mathbf{p},\cdot)$ always admits a minimum on $\mathcal{A}_0$;
- If $\Delta(\mathbf{p},\cdot)$ attains its minimum value on $\mathcal{A}_1$; then there is a unique coherent assessment $\mathbf{q}_{\underline{\alpha}}$ on $\mathcal{E}$ such that $\Delta(\mathbf{p},\underline{\boldsymbol{\alpha}})$ is minimum;
- If $\Delta(\mathbf{p},\cdot)$ attains its minimum value on $\mathcal{A}_0 \setminus \mathcal{A}_1$, then any distribution $\underline{\boldsymbol{\alpha}} \in \mathcal{A}_0$ that minimize $\Delta(\mathbf{p},\cdot)$ induce the same significant conditional probabilities $(\mathbf{q}_{\underline{\alpha}})_j$ on the conditional events $E_j|H_j$ such that $\underline{\alpha}(H_j) > 0$.
- Amongst the distributions $\underline{\boldsymbol{\alpha}} \in \mathcal{A}_0$ that minimize $\Delta(\mathbf{p},\cdot)$ there exists at least one $\widetilde{\boldsymbol{\alpha}}$ that maximize the number of positive conditioning events $\widetilde{\alpha}(H_j) > 0$ so that $\mathbf{q}_{\widetilde{\alpha}}$ has the largest number of uniquely determined components.

The last three items are the crucial ones: for precise numerical evaluations $\mathbf{p}$, they always guaranty the existence of a coherent assessment $(\mathcal{E},\mathbf{q}_{\widetilde{\alpha}})$ "close as much as possible" to $(\mathcal{E},\mathbf{p})$. And if there is the need to explore deeper "zero layers" (see again [9] for details about this delicate and crucial notion), the procedure to determine $(\mathcal{E},\mathbf{q}_{\widetilde{\alpha}})$ can be easily iterated over the residual conditioning events $E_i|H_i$ with $\widetilde{\alpha}(H_i) = 0$. A fully detailed procedure has been proposed in [5]. Note anyhow that the mixed-integer programs to determine $\widetilde{\boldsymbol{\alpha}}$

$$\max \sum_{E_i|H_i\in\mathcal{E}} I(\alpha'(H_i)) \quad (9)$$

s.t.

[1] Actually $\Delta(\mathbf{p},\boldsymbol{\alpha})$ turns out to be a generalization of the sum of two different "Bregman divergences".

$$\boldsymbol{\alpha}' \in \mathcal{A}_0 \tag{10}$$
$$\Delta(\mathbf{p}, \boldsymbol{\alpha}') = \Delta(\mathbf{p}, \underline{\boldsymbol{\alpha}}) \tag{11}$$
$$I(\alpha'(H_i)) = \begin{cases} 1 & \text{if} \quad \alpha'(H_i) > 0 \\ 0 & \text{if} \quad \alpha'(H_i) = 0 \end{cases} \tag{12}$$

can be operationally solved through the equivalent ordinary non-linear programs

$$\max \sum_{j=1}^{n} y_j \tag{13}$$
s.t.
$$0 \leq y_j \leq 1, \quad j = 1, \ldots, n \tag{14}$$
$$\frac{1}{m}\alpha'(H_j) \geq y_j, \quad j = 1, \ldots, n \tag{15}$$
$$m = \min_{H_i : \alpha'(H_i) > 0} \alpha'(H_i) \tag{16}$$
$$\boldsymbol{\alpha}' \in \mathcal{A}_0 \tag{17}$$
$$\Delta(\mathbf{p}, \boldsymbol{\alpha}') = \Delta(\mathbf{p}, \underline{\boldsymbol{\alpha}}). \tag{18}$$

3 Merging precise assessments with overlapping domains

Let us see how the aforementioned properties of $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ could help us on the merging problem. In particular we firstly focus one the specific case where the domains $\mathcal{E}_s$ overlaps, i.e. $\mathcal{E}_{s'} \bigcap \mathcal{E}_{s''} \neq \emptyset$, for some s' and s'' in S, and the numerical parts are all precise assessments (hence of the type $\mathbf{p}_s = (p1_s, \ldots, pn_s)$, $s \in S$).

This will imply that some of the conditional events $Ei|Hi \in \mathcal{E}$ of the joined support will be duplicated and with associated, possibly, different, precise values pi_s. For the sake of simplicity let us start by supposing that this will appear just for a single conditional event $Ed|Hd$ with just two distinct associated conditional probabilities $pd_{s'} \neq pd_{s''}$. The structure of $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ remains the same, but using its expression (7) we can join the two terms involving $Ed|Hd$ obtaining

$$\alpha(Hd)\left(q_d \ln(\frac{q_d^2}{pd_{s'}pd_{s''}}) + (1-q_d)\ln(\frac{(1-q_d)^2}{(1-pd_{s'})(1-pd_{s''})})\right). \tag{19}$$

Since, by hypothesis, the two distinct assessments $(\mathcal{E}, \mathbf{p}_{s'})$ and $(\mathcal{E}, \mathbf{p}_{s''})$ were coherent, the optimal solution $\widetilde{\boldsymbol{\alpha}}$ of the minimization problem of $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ under the constraint of $\boldsymbol{\alpha} \in \mathcal{A}_0$ will turn out as convex combination of the distributions compatible with $(\mathcal{E}, \mathbf{p}_{s'})$ and $(\mathcal{E}, \mathbf{p}_{s''})$. Consequently, the compromise value for q_d turns out to be simply the value in (0,1) minimizing (19), all the others terms in (7) being zero.

The same will not happen in general when there are more than one repeated conditional event in $\mathcal{E}$. In fact the set of coherent conditional assessment on $\mathcal{E}$ is not convex, hence it is not always guaranteed the optimal solution $\widetilde{\boldsymbol{\alpha}}$ being a convex combination of the distributions compatible with the single sources of information. Anyhow, the form of $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ remains technically unchanged and we have the same complexity to minimize it.

Let us see how this works with a simple example.

Example 1 *By borrowing the framework from [13], we consider two coincident domains* $\mathcal{E}_1 \equiv \mathcal{E}_2 = [C|A, C|B, C|A \vee B]$ *built by three basic unconditional logically independent events* A, B, C. *Hence the whole sample space would be of 8 atoms, but those inside* $H^0 \equiv A \vee B$ *are 6. The set of coherent assessments on* $\mathcal{E} = \mathcal{E}_1 = \mathcal{E}_2$ *is made by the triples* $[q_1, q_2, q_3] \in (0,1)^3$ *with the last component* q_3 *forced to belong to the range* $[\frac{q_1\, q_2}{q_1+q_2-q_1\, q_2}, \frac{q_1+q_2-2q_1\, q_2}{1-q_1\, q_2}]$ *(see Fig.1). Note the evident non-convexity of such coherent set.*

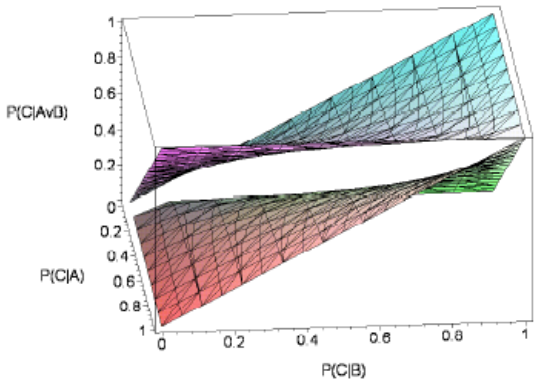


Figure 1: The lower and upper bounds for coherent assessments on $\mathcal{E} = [C|A, C|B, C|A \vee B]$

Let us consider two distinct coherent assessments

$\mathcal{E}_1 \equiv \mathcal{E}_2$	$C\|A$	$C\|B$	$C\|A \vee B$
$\mathbf{p}_1$	.2	.3	.14
$\mathbf{p}_2$	.2	.3	.4

. (20)

We can see that we have only the third element $C|A \vee B$ *with two distinct probability values. So the joint assessment results*

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$	$C\|A \vee B$
$\mathbf{p}$	.2	.3	.14	.4

. (21)

By performing the minimization[2] *of* $\Delta(\mathbf{p}, \boldsymbol{\alpha})$, *we get directly an optimal distribution* $\widetilde{\boldsymbol{\alpha}}$ *that makes positive all the conditioning events. So we obtain directly as merged assessment the following compromise* $\mathbf{q}_{\widetilde{\alpha}}$ *that leaves the first two assessment unchanged, while the third is a convex combination of the original ones:*

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$
$\mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$	.2	.3	.248

. (22)

If we consider a third assessment only on the first element $\mathbf{p}_3 \equiv P(C|A) = .3$ *we have to consider the joint assessment*

$\mathcal{E}$	$C\|A$	$C\|A$	$C\|B$	$C\|A \vee B$	$C\|A \vee B$
$\mathbf{p}$	.2	.3	.3	.14	.4

(23)

[2]The numerical results along all the examples have been obtained trough the nonlinear optimization software CONOPT of the package General Algebraic Modeling System (GAMS) [3]

that has as merged solution

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$
$\mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$	.245	.287	.254

. (24)

Note that also the value associated to $C|B$ has been changed, being incoherent to remain on its original value.

In the joint assessment (21) we miss the information that some element of the joint domain has different frequency. This suggest us a further generalization. In fact, it is possible to associate different weights to the elements of the joined assessment $(\mathcal{E}, \mathbf{p})$. This to reflect either possible repetitions of the values or different trust on the various sources of information. We can denote by $\mathbf{w} = [w_1, \ldots, w_n]$ such weights and adjust the expression of $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ as

$$\Delta^{\mathbf{w}}(\mathbf{p}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha(Hi) \left(q_i \ln(\frac{q_i^{w_i}}{p_i^{w_i}}) + (1 - q_i) \ln(\frac{(1-q_i)^{w_i}}{(1-p_i)^{w_i}}) \right). \quad (25)$$

Let us see what are its effects on the previous example

Example 2 *If we consider on the joint assessment (23) how many times each value has been assessed, we have the following weights association*

$\mathcal{E}$	$C\|A$	$C\|A$	$C\|B$	$C\|A \vee B$	$C\|A \vee B$
$\mathbf{p}$	.2	.3	.3	.14	.4
$\mathbf{w}$	2	1	2	1	1

. (26)

By performing the minimization of (25) we obtain the following merged solution

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$
$\mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$	.229	.292	.255

. (27)

By comparing (24) with (27), we can note that the highest weights have the effect to "attract" the compromise solution to the associated values.

4 Merging knowledge bases with disjoint supports

Let us face now a different situation. Specifically, consider the case where the supports $\mathcal{E}_s$ of the various sources are disjoint. Equivalently, consider when we have all multiplicity 1 in $\mathcal{E}$. This is the case for example when we join together different knowledge bases. Of course the interesting cases are those where, even disjoint, the supports $\mathcal{E}_s$ are correlated.

The problem of finding a compromise among the different assessments obviously arise when the joined assessment $(\mathcal{E}, \mathbf{p})$ turns out to be incoherent.

If all the numerical parts are precise, then the solution is simply the assessment $\mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$ derived by the optimal solution $\widetilde{\boldsymbol{\alpha}}$ that minimize $\Delta(\mathbf{p}, \boldsymbol{\alpha})$. A different approach is instead needed in presence of some imprecise assessment. The solution will be the result of several corrections, each one obtained by fixing one

of the boundary values (lbi_s or ubi_s) and letting the other components to vary inside their ranges $[lbj_s, ubj_s]$ (that could be actually points pj_s).

A detailed description of the procedure has been given in [6]. Here we give an idea of how it could work by illustrating a simplified example.

Example 3 *Let us suppose to merge a physician expertise about the separate validity of two distinct bio-markers for a specific tumoral lesion. Specifically, let D denote the event that a patient being diagnosed to have the lesion, F the expression of the first bio-marker and S the expression of the second one. There are not particular logical relations among the three unconditional events D, F and S.*

Consider the following physician opinion about the ratio of patients with specific symptoms being diagnosed to have the lesion and about the expression of the two bio-markers among those patients:

$\mathcal{E}_1$	D	$F\|D$	$S\|D$
$\mathbf{p}_1$	.51	.231	.77

. (28)

On the other hand, the physician is interested in the correlation between the two bio-markers. He founds out in the literature the expected ratio of expressions of the first bio-marker among those that express also the second between 65% *and* 75%*. Hence we have a second source of information composed by a single imprecise assessment*

$\mathcal{E}_2$	$F\|S$
$\mathbf{p}_2$	$[.65, .75]$

. (29)

Note that the joint assessment

$\mathcal{E}$	D	$F\|D$	$S\|D$	$F\|S$
$\mathbf{p}$	.51	.231	.77	$[.65, .75]$

(30)

seen as a lower-upper conditional probability assessment is g-coherent [1] (a notion equivalent to Walley's avoiding uniform loss). In fact there exists at least one coherent (in de Finetti's sense) precise conditional probability assessment compatible with the numerical ranges. On the other side, it is not coherent in the more strict sense (see for example [8]) because, by taking $\mathbf{p}_1$ as valid, not all the values inside the range [.65,.75] for $P(F|S)$ can be reached. In particular the upper bound .75 is outside the coherent extension $[.0005, .6886]$ *for $P(F|S)$.*

If we try to adjust the upper evaluation

$\mathcal{E}$	D	$F\|D$	$S\|D$	$F\|S$
$\bar{\mathbf{p}}$	.51	.231	.77	.75

(31)

we get as closest coherent assessment the following:

$\mathcal{E}$	D	$F\|D$	$S\|D$	$F\|S$
$\mathbf{q}$	.49	.256	.752	.723

. (32)

Hence we can conclude that the best coherent merging compromise is

$\mathcal{E}$	D	$F\|D$	$S\|D$	$F\|S$
$\mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$	$[.49, .51]$,	$[.231, .256]$	$[.752, .77]$	$[.65, .723]$

. (33)

Note that even being $\mathbf{p}_1$ *a precise assessment, in the merging solution we have all interval values. This sounds reasonable because the information* $\mathbf{p}_1$ *provided by the physician is usually based on his specific knowledge base, while the literature information* $\mathbf{p}_2$ *is usually made by collecting several case studies and the merging procedure reflect this.*

5 Conclusion

In this paper we have shown that the tool of the discrepancy measure $\Delta(\mathbf{p}, \boldsymbol{\alpha})$, introduced originally to adjust incoherent partial conditional probability assessments, turns out useful also for reasonable merging of different source of information.

This procedure does not require any particular further theoretical investigation but just a careful use of the already stated properties of the discrepancy measure.

Even being limited to specific situations, the merging procedures reported here are quite representative of the most common applications. Obviously a much deep and complete analysis is needed. In particular, formal properties (axioms) that our aggregation operation satisfies must be investigated. We can anticipate that surely it is not associative, as already noted in [4], while it is surely symmetric and *preserves unanimity* (see [22]).

References

[1] V. Biazzo and A. Gilio: A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments. *Int. J. of Approximate Reasoning*, 24, 251-272, (2000).

[2] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, 7:200–217, (1967).

[3] A. Brooke , D. Kendrick , A. Meeraus and R. Raman. GAMS: a Users Guide (2003), Washington, D.C.: GAMS Development Corp.

[4] A. Capotorti, G. Regoli. Coherent correction of inconsistent conditional probability assessments. in *Proc. of IPMU'08 - Malaga (Es)*, (2008).

[5] A. Capotorti, G. Regoli, F. Vattari. Theoretical properties of a discrepancy measure among partial conditional probability assessments. *manuscript submitted to IJAR for acceptance*, (2009).

[6] A. Capotorti, G. Regoli, F. Vattari. On the use of a new discrepancy measure to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities. *Proceedings of* ISIPTA'09 *int. conf.*. (2009).

[7] G. Coletti: Coherent numerical and Ordinal probabilistic assessments. *IEEE Transaction on Systems, Man, and Cybernetics*, **24**, 1747-1754 (1994).

[8] G. Coletti, R. Scozzafava. The role of coherence in eliciting and handling imprecise probabilities and its application to medical diagnosis. *Information Science*, 13:41–65 (2000).

[9] G. Coletti, R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, Series "Trends in Logic", (2002).

[10] R.M. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press. (1991).

[11] G. de Cooman, M. C. M. Troffaes. Coherent lower previsions in systems modelling: products and aggregation rules. *Reliability Engineering and System Safety*, 85(1-3):113-134, (2004).

[12] C. Genest, JV Zidek. Combining probability distributions: A critique and an annotated bibliography, *Statistical Science*, 1:114–148, (1986).

[13] A. Gilio. Probabilistic Relations Among Logically Dependent Conditional Events, *Soft Computing*,3:154–161, (1999).

[14] S. Holzer. On coherence and conditional prevision. *Bull. Unione Matematica Italiana, Analisi funzionale e applicazioni*. 6(4): 441-460, (1985).

[15] G. Kern-Isberner, W. Rödder. Belief revision and information fusion on optimum entropy. *Int. J. of Intelligent Systems*, 19, 837–857 (2004).

[16] O. Kriz, Conditional problem for Objective Probability, *Kybernetica*, 34(1), 27–40, (1998).

[17] F. Lad, *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*, New York: John Wiley, (1996).

[18] S. Moral, J. del Sagrado. Aggregation of imprecise probabilities. In: Bouchon-Meunier, *Aggregation and Fusion of Imperfect Information*, Physica-Verlag, New York, 162–188, (1998).

[19] R. F. Nau. The aggregation of imprecise probabilities, *Journal of Statistical Planning and Inference* 105(1):265–282, (2002).

[20] R. Nau, V. R. Jose and R. Winkler: Scoring Rules, Entropy, and Imprecise Probabilities, in *Proc. 5th Int. Symp. on Imprecise Probability : Theorie and Applications. ISIPTA'07 - Prague (CZ)*, 307–315, (2007).

[21] S. Rässler, *Statistical Matching: a frequentist theory, practical applications and alternative Bayesian applications*, Springer, (2002).

[22] M.C.M. Troffaes Generalizing the Conjunction Rule for aggregating Conflict Expert Opinions *International Journal of Intelligent Systems*, 21: 361–380, (2006).

[23] B. Vantaggi. Statistical matching of multiple sources: A look through coherence, *International Journal of Approximate Reasoning*, 49(3):701–711, (2008).

[24] P. Walley. *Statistical reasoning with Imprecise Probabilities*, Chapman and Hall, London, (1991).

Belief Conditioning Rules for Classic Belief Functions*

Milan Daniel
Institute of Computer Science
Academy of Sciences of the Czech Republic
milan.daniel@cs.cas.cz

Abstract

The classic belief conditioning rules (BCRs) and DSm BCRs applied to classic belief functions are briefly recalled in the contribution. A general idea of belief conditioning by a given evidence is analysed and a new plain BCR is presented. This rule is compared with both the classic and the DSm BCRs. Finally, general formulas for described BCRs and a new general BCR are presented and defined.

1 Introduction

Belief functions are one of the widely used formalisms for uncertainty representation and processing. Belief functions enable representation of incomplete and uncertain knowledge, belief updating and combination of evidence. Originally belief functions were introduced as a principal notion of *Dempster-Shafer Theory* (DST) or the Mathematical Theory of Evidence [8, 12].

For combination of beliefs, Dempster's rule of combination and a series of its alternatives is used in DST. Namely *Dempster's rule of conditioning* [8] is used for belief function conditioning by evidence, i.e. in the case that we obtain a sure assupmtion that the true is definitely in some proper subset of the frame of discernment. The alternative *belief focusing rule* completely ignores basic belief mases of all focal elements which are not a subset of the conditioning set.

In a new DSm (Dezert-Smarandache) approach to belief functions [5] a long series of 31 *belief conditioning rules* (BCRs) was defined [9]. Due to the fact that the DSm approach can be considered both as a generalization and also a special case of Dempster-Shafer theory [3], the series of DSm BCRs is reduced to 9 conditioning rules for classic belief functions (where one of them is belief focusing in fact). All BCRs (including Dempster's rule) add some additional information to the conditioned belief functions. Unfortunately majority of DSm BCRs add more information than it is necessary. Some of these rules are non-intuitive or even counter-intuitive, thus only BCR12 is considered to be reasonably useful [4]. Nevertheless the analysis of the entire series of 31 DSm BCRs motivated

*This work was supported by ESF EUROCORES FP006 project ICC/08/E018 of the Grant Agency of the Czech Republic, and in part by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

the author to look for a conditioning rule which adds no additional information within conditioning.

Refusing the usual assumption that conditioning of a Bayesian belief function should be again Bayesian belief function, we can define a new *plain rule of belief conditioning*, which adds no additional information to the conditioned belief. By application of a probability transformation to the resulting conditioned belief function we obtain a Bayesian belief function again, thus our reduction of assumptions brings no principal limitation either on the belief level or on the decisional (pignistic) level.

On the other hand, when applying various probability transformations [1], we obtain various conditioned Bayesian belief functions; in the special case of normalised belief of singletons, we obtain the same conditioned Bayesian belief function as when the belief focusing, BCR12 or Dempster's rule of conditioning are used.

Properties of the plain BCR are described and presented in examples. Multiple conditioning with the plain BCR is weakly commutative and associative. The plain BCR can be expressed using both Dubois-Prade [6] and Yager's [11] rules of combination of belief functions.

In the end, a new general formula of BCR is presented, which expresses not only all of the above mentioned rules (belief focusing, BCR12, Dempster's rule of conditioning, the plain BCR), but also a new general conditioning rule — a combination of the four previously presented ones.

2 Preliminaries

Let us assume an exhaustive finite *frame of discernment* $\Omega = \{\omega_1, ..., \omega_n\}$, whose elements are mutually exclusive.

A *basic belief assignment (bba)* is a mapping $m : \mathcal{P}(\Omega) \longrightarrow [0,1]$, such that $\sum_{A\subseteq\Omega} m(A) = 1$, the values of bba are called *basic belief masses (bbm).*[1] $\mathcal{P}(\Omega) = \{X | X \subseteq \Omega\}$ is often denoted also by 2^Ω.

A *belief function (BF)* is a mapping $Bel : \mathcal{P}(\Omega) \longrightarrow [0,1]$, $Bel(A) = \sum_{\emptyset\neq X\subseteq A} m(X)$. A *plausibility function* is a mapping $Pl : \mathcal{P}(\Omega) \longrightarrow [0,1]$, $Pl(A) = \sum_{\emptyset\neq X\cap A} m(X)$. Belief function Bel, Plausibility function Pl and the corresponding bba m uniquely correspond to each other.

Dempster's (conjunctive) rule of combination $\oplus$ is given as $(m_1 \oplus m_2)(A) = \sum_{X\cap Y=A} K m_1(X)m_2(Y)$ for $A \neq \emptyset$, where $K = \frac{1}{1-\kappa}$, $\kappa = \sum_{X\cap Y=\emptyset} m_1(X)m_2(Y)$, and $(m_1 \oplus m_2)(\emptyset) = 0$, see [8]; putting $K = 1$ and $(m_1 \oplus m_2)(\emptyset) = \kappa$ we obtain the *non-normalized conjunctive rule of combination* $\textcircled{\cap}$.

Yager's rule of combination $\textcircled{Y}$, see [11], is given as $(m_1 \textcircled{Y} m_2)(\emptyset) = 0$, $(m_1 \textcircled{Y} m_2)(A) = \sum_{X,Y\subseteq\Theta,\, X\cap Y=A} m_1(X)m_2(Y)$ for $\emptyset \neq A \subset \Theta$, and $(m_1 \textcircled{Y} m_2)(\Theta) = m_1(\Theta)m_2(\Theta) + \sum_{X,Y\subseteq\Theta,\, X\cap Y=\emptyset} m_1(X)m_2(Y)$;

Dubois-Prade's rule of combination $\textcircled{DP}$ is given as $(m_1 \textcircled{DP} m_2)(A) = \sum_{X,Y\subseteq\Theta,\, X\cap Y=A} m_1(X)m_2(Y) + \sum_{X,Y\subseteq\Theta,\, X\cap Y=\emptyset, X\cup Y=A} m_1(X)m_2(Y)$ for $\emptyset \neq A \subseteq \Theta$, and $(m_1 \textcircled{DP} m_2)(\emptyset) = 0$, see [6].

Probabilistic transformations: pignistic transformation $BetT$, (normalized) plausibility transformation Pl_T, (normalized) belief transformation Bel_T, pro-

[1] $m(\emptyset) = 0$ is often assumed in accordance with Shafer's definition [8]. A classical counter example is Smets' Transferable Belief Model (TBM) which admits $m(\emptyset) \geq 0$ [10].

portional belief transformation $Prop_{Bel}T$, are mappings from the set of belief functions to the set of probability functions, see [1, 2].

In this contribution we deal with belief conditioning by event, i.e. we suppose a sure assumption, that the truth (the true element $\omega_0 \in \Omega$) is in some specified proper subset A of frame of discernment Ω.

3 Classic belief conditioning rules

There are several equivalent forms of *Dempster's rule of conditioning (DRC)*. The original introduced by Shafer in [8] uses plausibility measure: $Pl(X|A) = \frac{Pl(X \cap A)}{Pl(A)}$, the expression which uses bba is the following

$$m(X|A) = \tfrac{1}{1-k} \textstyle\sum_{Y \cap A = X} m(Y),$$

for $X \subseteq A$, where $k = \sum_{Y \cap A = \emptyset} m(Y)$; $m(X|A) = 0$ for $X \not\subseteq A$. The rule is defined (applicable) whenever $Pl(A) > 0$, i.e., whenever there exists some $Y \cap A \neq \emptyset$ such that $m(Y) > 0$. For a comparison with DSm BCRs, we can equivalently write:

$$m(X|A) = m(X) + \sum_{\substack{Y \cap A = X \\ Y \neq X}} m(Y) + \frac{\sum_{Y \cap A = X} m(Y)}{\sum_{Y \cap A \neq \emptyset} m(Y)} \sum_{Y \cap A = \emptyset} m(Y)$$

For DRC it holds that $m(X|A) = (m \oplus m_A)(X)$, where $m_A(A) = 1$, $m_A(X) = 0$ for $X \neq A$.

There is another belief conditioning rule, called also *belief focusing*[2] *(BFR)*:

$$m(X||A) = \tfrac{m(X)}{Bel(A)} = \tfrac{m(X)}{\sum_{Y \subseteq A} m(Y)}$$

for $X \subseteq A$, $m(X||A) = 0$ for $X \not\subseteq A$. This rule is applicable whenever $Bel(A) > 0$, i.e., whenever there exists some $\emptyset \neq Y \subseteq A$ such that $m(Y) > 0$, see [7].

4 DSm belief conditioning rules

There are 31 DSm belief conditioning rules (BCRs) defined and presented in [9]. These rules are defined on DSm hyper-power sets constructed from frames of with overlapping elements. Considering the classic case with exclusive elements, some of these rules become mutually equivalent. Hence we obtain 9 different DSm BCRs for classic belief functions which are defined on $\mathcal{P}(\Omega)$.

BCR1 is equivalent to the belief focusing rule for classic BFs. Unfortunately, majority of DSm BCRs are not intuitive and some of them are even counter-intuitive. All these rules add additional information to conditioned BFs within conditioning process. Only BCR12, which adds the least amount of additional information (and which is equivalent to BCR2 for classic BFs), has been founded to be reasonable for belief conditioning in [4]. Thus we consider only BCR12 in this text.

[2] We use a notation $m(\_||\_)$ to distinguish it from Dempster's conditioning rule $m(\_|\_)$. This rule was mentioned in [7], unfortunately, the authors of that chapter does not know its original publication.

We can present BCR12 for classic belief functions, i.e. for BFs on Shafer's model in DSm terminology, as it follows[3] (see [4]):

$$m_{BCR12}(X \amalg A) = \sum_{W \cap A = X} m(W) \; + \; \frac{m(X)}{Bel(A)} \cdot \sum_{Z \cap A = \emptyset} m(Z)$$

for $X \subseteq A$, $m_{BCR12}(X \amalg A) = 0$ for $X \not\subseteq A$. This rule is applicable[4] whenever $Bel(A) > 0$ or $Pl(A) = 1$.

5 General idea of belief conditioning rules

Let us look at particular BCRs, how bbms of individual focal elements are transformed from an input belief function to the corresponding conditioned one. We can divide focal elements of input BFs according to their relation to conditioning set A into 3 disjoint subsets: focal elements which are subset of the conditioning set A: $S = \{X_S | X_S \subseteq A \subseteq \Omega\}$, elements which are not subset of A but which have non-emty intersection with A: $I = \{X_I | X_I \subseteq \Omega, X_I \not\subseteq A, X_I \cap A \neq \emptyset\}$, and the remainig focal elements which are disjunctive with A: $D = \{X_D | X_D \subseteq \Omega, X_D \cap A = \emptyset\}$, see figure 1. We can use $S_{\Omega,A}, I_{\Omega,A}, D_{\Omega,A}$ for specification of the frame of discernment Ω and the conditioning set A. In the same way we can split entire $\mathcal{P}(\Omega)$ into 3 disjoint parts $S_{\Omega,A} \cup I_{\Omega,A} \cup D_{\Omega,A} = \mathcal{P}(\Omega)$.

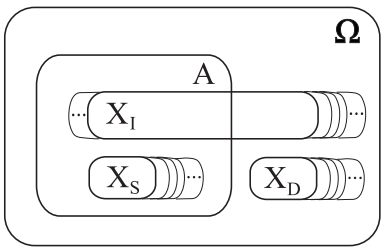


Figure 1: Three disjoint subsets of focal elements.

BFR performs normalization of bbms of focal elements from S (briefly bbms of X_Ss). In another words, we can say that BFR keeps bbms of all X_Ss and proportionalizes bbms of all focal elements from I and D (bbms of X_Is and X_Ds) according to bbms of X_Ss.
BCR12 also keeps bbms of X_Ss, it transfers bbms of X_Is to $X_I \cap A$ (adds $m(X_I)$ to $m(X_I \cap A)$ and finally proportionalizes bbms of X_Ds according to input bbms of X_Ss.
DRC keeps again bbms of X_Ss, and it transfers bbms of X_Is to $X_I \cap A$ in the same way as BCR12. But the final proportionalization of bbms of X_Ds is performed according to bbms of X_Ss increased by bbms of corresponding X_Is (such that $X_S = X_I \cap A$).

[3] We use a notation $m(\_ \amalg \_)$ to simply distinguish bba m conditioned by BRC12 from the same bba m conditioned by Dempster's conditioning rule $m(\_ | \_)$ and from that conditioned by the belief focusing $m(\_ || \_)$.

[4] We have to mention here the Dezert-Smarandache idea of extension BCR definition domains, $m_{BCRi}(A \amalg A) = 1$, whenever BCRi is not defined; for comments about correctness of application of this idea to DSm BCRs see [4].

Have a look at this in general: all X_Ss are completely contained in A, they are in full accord with A, thus there is no need to transfer or redistribute their bbms within the conditioning process. This is fulfilled by all 3 BCRs presented above (and this is true also for all BCRs from [9]).

In the case of X_Is we have an input belief that ω_0 (true) is in X_I and sure assumption/condition that $\omega_0 \in A$, thus we should believe that ω_0 is in $X_I \cap A \neq \emptyset$ under the condition. Thus, there really should be no problem with X_Is. This is fullfilled by BCR12 and DRC but not by BFR (also not by any BCRs from [9] which are not equivalent to BCR12 in the classic Shafer's case). BFR completely ignores bbms of all X_Is and it adds to the input BF the additional assumption that the conditioned bbms must be in the same ratio as the input bbms of X_Ss are.

The most complicated is the situation of X_Ds: similarly to the previous case, we have an input belief that ω_0 (true) is in X_D and sure assumption/condition that $\omega_0 \in A$, but $X_D \cap A = \emptyset$. How to solve this conflicting situation? We have the sure assumption/condition that ω_0 must be in A and we have a belief that it is in X_D, i.e. out of A. We assume, that we have no other additional belief, information or assumption within the conditioning process. Thus there is no reason for any proportionalization of bbms of X_Ds, there is neither reason for any other redistribution of bbms of X_Ds among proper subsets of A. We only assume that ω_0 is in A thus we have to transfer bbms of X_Ds to bbm of A (to add all $m(X_D)$ to $m(A)$). This is performed by none of the above BCRs (nor by any BCRs from [9]). Hence a new BCR should be defined as it follows in the next section.

6 The plain belief conditioning rule

Based on the ideas from the previous section, we can define a new *plain belief conditioning rule* (plain BCR) as it follows:

$$m(X\vdots A) = \sum_{Y \cap A = X} m(Y)$$

for $X \subset A$,

$$m(A\vdots A) = \sum_{Y \cap A = A} m(Y) + \sum_{Y \cap A = \emptyset} m(Y),$$

$m(X\vdots A) = 0$ for $X \not\subseteq A$. In difference from the other BCRs, the plain BCR is defined for any BF defined on $\mathcal{P}(\Omega)$. For a comparison with DSm BCRs, we can equivalently write: $m(X\vdots A) = m(X) + \sum_{\substack{Y \cap A = X \\ Y \neq X}} m(Y)$ for $X \subset A$, and $m(A\vdots A) = m(A) + \sum_{\substack{Y \cap A = A \\ Y \neq X}} m(Y) + \sum_{Y \cap A = \emptyset} m(Y)$.

6.1 Properties of the plain BCR

When a Bayesian BF is conditioned by any of BFR, BCR12 and DRC, the resulting conditioned BF is again Bayesian. This does not hold in the case of the plain BCR as bbms of all singletons out of the conditioning set A are summed to the $m(A)$. Thus Bayesianity is kept only in non-interesting trivial cases by the plain BCR: 1) for singleton A and 2) for BBFs with positive bbms only for singletons from A.

Thus Shafer's assumption of coincidence of conditioning of BBFs with probability conditioning is not satisfied by the plain BCR. Is it a serious disadvantage for BCR? It is not. When working with BFs we should use their wider expressibility than probability distributions have, and there is no necessity to be forced to Bayesian results in the credal level. The result is more general as it is possible to use different probability transformations and to obtain different corresponding results on the decisional level. Moreover when (normalized) belief probability transformation Bel_T is applied to a BBF conditioned by the plain BCR, we obtain just the same result as if Dempster's rule of conditioning is applied to the input BBF.

Let us suppose situation, where two or more conditioning sets subsequently appear now. Thus conditioning rule should be applied twice or several times to the given belief function(s). As we want accept all the subsequentially appeared conditioning sets A_i, the resulting conditioned BF may have positive bbas only for focal elements which are subset or equal to the intersection $\bigcap_i A_i$ of all the considered conditioning sets.

Statement 1 *The plain belief conditioning rule is weakly commutative in the following sense:* $m((X\dot{:}A)\dot{:}A \cap B) = m((X\dot{:}B)\dot{:}A \cap B)$.

This statement directly follows the fact that the following holds true: $m((X\dot{:}A)\dot{:}A \cap B) = m(X\dot{:}A \cap B)$.

Corollary 2 *The plain belief conditioning rule is weakly associative in the following sense:* $m(((X\dot{:}A)\dot{:}A \cap B)\dot{:}A \cap B \cap C) = m(((X\dot{:}C)\dot{:}B \cap C)\dot{:}A \cap B \cap C)$.

A relation of the plain BCR and of the belief combination rules is expressed by the following statements.

Statement 3 *(i) For the plain belief conditioning rule and for Yager's rules of combination the following holds true:*

$$m(X\dot{:}A) = (m\textcircled{Y}m_A)(X),$$
$$m(A\dot{:}A) = (m\textcircled{Y}m_A)(A) + (m\textcircled{Y}m_A)(\Omega)$$
$$m(\Omega\dot{:}A) = 0,$$

where $X \neq A, \Omega$, $m_A(A) = 1,\ m_A(Y) = 0$ *for* $Y \neq A$.
(ii) For the plain belief conditioning rule and for Dubois-Prade rules of combination the following holds true:

$$m(X\dot{:}A) = (m\textcircled{D}m_A)(X),$$
$$m(A\dot{:}A) = \sum_{A \subseteq Y} (m\textcircled{D}m_A)(Y)$$
$$m(Z\dot{:}A) = 0,$$

where $X \subset A$, $Z \not\subseteq A$, $m_A(A) = 1,\ m_A(Y) = 0$ *for* $Y \neq A$.

We can summarize this by the following theorem.

Theorem 4 *For the plain belief conditioning rule and for Yager's and Dubois-Prade rules of combination the following holds true:*

$$m(X^{\cdot}A) = (m \textcircled{\cdot} m_A)(X) = (m \textcircled{@} m_A)(X),$$
$$m(A^{\cdot}A) = (m \textcircled{\cdot} m_A)(A) + (m \textcircled{\cdot} m_A)(\Omega) = \sum_{A \subseteq Y} (m \textcircled{@} m_A)(Y),$$

where $X \subset A$, $m_A(A) = 1,\ m_A(Y) = 0$ *for* $Y \neq A$,

$m(X^{\cdot}A) = 0$ *for* $X \not\subseteq A$.

It also holds that

$$m(X^{\cdot}A) = (m \textcircled{\cdot} m_A)(X^{\cdot}A) = (m \textcircled{@} m_A)(X^{\cdot}A),$$
$$m(X^{\cdot}A) = (m \textcircled{\cdot} m_A)(X|A) = (m \textcircled{@} m_A)(X|A),$$
$$m(X^{\cdot}A) = (m \textcircled{\cdot} m_A)(X \shortparallel A) = (m \textcircled{@} m_A)(X \shortparallel A).$$

For proper subsets of the conditioning set, the bbms of BF conditioned by the plain belief conditioning rule coincide with values obtained by Dubois-Prade and Yager's rules of combination with one argument fixed to categorical BF m_A. Value $m(A^{\cdot}A)$ is equal to the remainder of the sum of $m(X^{\cdot}A)$ to 1, where $X \subset A$. Thus the plain BCR is compatible with both Dubois-Prade and Yager's rules of combination.

7 Properties and comparison of BCRs

7.1 Definition Domains

BFR has the least definition domain $\mathcal{D}om_{BFR} = \{Bel \mid Bel(A) \neq \emptyset\}$ among the presented rules, $\mathcal{D}om_{BCR12} = \{Bel \mid Bel(A) \neq \emptyset \vee Pl(A) = 1\}$[5], $\mathcal{D}om_{DRC} = \{Bel \mid Pl(A) \neq \emptyset\}$, the plain BCR has the largest possible definition domain, i.e. set of all belief functions $\mathcal{D}om_{PBCR} = \{Bel\}$.

7.2 Conditioning of Bayesian belief functions

When a Bayesian BF is conditioned by any of BFR, BCR12 and DRC, the resulting conditioned BF is again Bayesian. In the case of the plain BCR the resulting conditioned BF is more general, usually not Bayesian.

7.3 Multiple conditioning

Let us suppose again situations, where two or more conditioning sets subsequently appear, thus conditioning rules should be applied twice or several times to the given belief function(s).

Both BFR and DRC are commutative and associative in the following sense: $m((X^{\cdot}A)^{\cdot}B) = m((X^{\cdot}B)^{\cdot}A)$, $m(((X^{\cdot}A)^{\cdot}B)^{\cdot}C) = m(((X^{\cdot}C)^{\cdot}B)^{\cdot}A)$. The plain BCR has only weak version of these properties, see the previous section. For BCR12 neither the weak version of these properties holds true, see the following example.

[5] We have to note, that Dezert & Smarandache additionally extended definition domains of all DSm BCRs with formula $m_{BCRi}(A|A) = 1$, for all BFs out of the original definition domains of the rules. Their idea is discussed in the Appendix of [4].

Example: Let suppose $\Omega_6 = \{\omega_1, \omega_2, ..., \omega_6\}$, conditioning sets $A = \{\omega_2, \omega_3, \omega_4\}$, $B = \{\omega_1, \omega_2, \omega_3\}$, and Bel given by m as it follows: $m(\{\omega_2\}) = 0.1$, $m(\{\omega_5, \omega_6\}) = 0.3$, $m(\{\omega_3, \omega_4, \omega_5\}) = 0.5$, $m(\{\omega_4, \omega_5, \omega_6\}) = 0.1$.
Thus we obtain $m(\{\omega_2\}\dot{:}A) = 0.1$, $m(\{\omega_3, \omega_4\}\dot{:}A) = 0.5$, $m(\{\omega_4\}\dot{:}A) = 0.1$, $m(A\dot{:}A) = 0.3$ and $m((\{\omega_2\}\dot{:}A)\dot{:}A \cap B) = 0.1$, $m((\{\omega_3\}\dot{:}A)\dot{:}A \cap B) = 0.5$, $m((A \cap B\dot{:}A)\dot{:}A \cap B) = 0.4$;
$m(\{\omega_2\}\dot{:}B) = 0.1$, $m(\{\omega_3\}\dot{:}B) = 0.5$, $m(B\dot{:}B) = 0.4$ and $m((\{\omega_2\}\dot{:}B)\dot{:}A \cap B) = 0.1$, $m((\{\omega_3\}\dot{:}B)\dot{:}A \cap B) = 0.5$, $m((A \cap B\dot{:}B)\dot{:}A \cap B) = 0.4$.
We further obtain $m(\{\omega_2\} \amalg A) = 0.15$, $m(\{\omega_3, \omega_4\} \amalg A) = 0.75$, $m(\{\omega_4\} \amalg A) = 0.1$, and $m((\{\omega_2\} \amalg A) \amalg A \cap B) = 0.25$, $m((\{\omega_3\} \amalg A) \amalg A \cap B) = 0.75$;
whereas $m(\{\omega_2\} \amalg B) = 0.5$, $m(\{\omega_3\} \amalg B) = 0.5$, and $m((\{\omega_2\} \amalg B) \amalg A \cap B) = 0.5$, $m((\{\omega_3\} \amalg B) \amalg A \cap B) = 0.5$.

7.4 Relation of BCRs to combination rules

DRC coincides with Dempster's rule of combination with one argument fixed to Bel_A, where corresponding bba is given as $m_A(A) = 1$, $m_A(X) = 0$ otherwise. The plain BCR compatible with Yager's and Dubois-Prade rules of combination with one argument fixed to Bel_A.

BFR and BCR12 are not compatible with any combination rule in this sense.

7.5 Addition of additional information during conditioning

The plain BCR adds no additional information within the conditioning process. DRC and BCR12 add additional information when bbas of conflicting focal elements (focal element from the set D, see Sect. 5) are normalized or proportionalized. BFR adds the greatest amount of additional information because bbas of focal elements intersecting conditioning set A are also normalized.

7.6 Comparison of BCRs

Let us compare the presented BCRs according to their above mentioned properties. We left aside compatibility of BCRs with combination rules. This property is criticised and strictly rejected by Dezert and Smarandache in the case of DRC and Dempster's rule of combination. On the other hand, this property does not look important for evaluation of rules, this property is not important for conditioning process, but it is important for understanding of nature of the conditioning rules, it is important for their compatibility with belief combination.

Both positive properties of BFR processing of BBFs and its commutativity and associativity are possesed also by DRC. On the other hand DRC has larger definition domain and adds less additional information, thus DRC is better than BFR is.

When comparing DRC and BCR12, both the rules keep Bayesians BFs and add the same amount of (different) additional information. DRC has greater definition domain and posses commutativity and associativity when multiple

conditioning is applied. Thus DRC is also better[6] than BCR12 is.

When comparing BCR12 and BFR, BCR12 may possible be evaluated as a better one from the applicational (def. domain and additional information), but not from the theoretical point of view (not even weak commutativity).

When comparing DRC and the plain BCR, DRC is commutative and associative whereas the plain BCR posses only weak version of these properties, and DRC keeps BBFs. Both the conditioning rules are compatible with some belief combination rule(s). The plain BCR is better from the applicational point of view (greater definition domain and less additional information). Hence it is not possible to say which of the rules is better in general.

When comparing the plain BCR with BCR12, the only advantage of BCR12 is keeping of BBFs, nevertheless it is not possible to say which of the rules is better in general.

When comparing all four presented BCRs, we can point out DRC and the plain BCR: as DRC is better than BFR and BCR12 and the plain BCR is not better or worse than the others in general. Moreover both DRC and the plain BCR are compatible with some combination rule: DRC with Dempster's one and the plain BCR with Yager's and Dubois-Prade rules of combination.

8 General formula for belief conditioning rules

We can reformulate BFR, BCR12, DRC, and the plain BCR on their definition domains as it follows:

$m(X||A) = m(X)\frac{\sum_{Y\cap A\neq\emptyset} m(Y)}{\sum_{Y\subseteq A} m(Y)} + m(X)\frac{\sum_{Y\cap A=\emptyset} m(Y)}{\sum_{Y\subseteq A} m(Y)}$ for $X \subseteq A$,

$m_{BCR12}(X \shortparallel A) = \sum_{W\cap A=X} m(W) \; + \; \frac{m(X)}{Bel(A)} \cdot \sum_{Z\cap A=\emptyset} m(Z)$ for $X \subseteq A$,

$m(X|A) = \sum_{Y\cap A=X} m(Y) \; + \; \frac{\sum_{Y\cap A=X} m(Y)}{\sum_{Y\cap A\neq\emptyset} m(Y)} \sum_{Y\cap A=\emptyset} m(Y)$ for $X \subseteq A$,

$m(X\dot{:}A) = \sum_{Y\cap A=X} m(Y)$ for $X \subset A$, $m(A\dot{:}A) = \sum_{Y\cap A=A} m(Y) + \sum_{Y\cap A=\emptyset} m(Y)$,

$m(X||A) = m_{BCR12}(X \shortparallel A) = m(X|A) = m(X\dot{:}A) = 0$ for $X \not\subseteq A$.

Using the idea $m(X \wr A) = p\, m(X||A) + q\, m_{BCR12}(X \shortparallel A) + r\, m(X|A) + s\, m(X\dot{:}A)$, where $p, q, r, s \in \{0, 1\}$ such that $p + q + r + s = 1$, we obtain the following general formulas for the presented belief conditioning rules:

$$m(X \wr A) = p\, m(X)\frac{\sum_{Y\cap A\neq\emptyset} m(Y)}{\sum_{Y\subseteq A} m(Y)} + (q + r + s) \sum_{Y\cap A=X} m(Y)$$

$$+(p + q)m(X)\frac{\sum_{Y\cap A=\emptyset} m(Y)}{\sum_{Y\subseteq A} m(Y)} + r\frac{\sum_{Y\cap A=X} m(Y)}{\sum_{Y\cap A\neq\emptyset} m(Y)} \sum_{Y\cap A=\emptyset} m(Y),$$

$$m(A \wr A) = p\, m(A)\frac{\sum_{Y\cap A\neq\emptyset} m(Y)}{\sum_{Y\subseteq A} m(Y)} + (q + r + s) \sum_{Y\cap A=A} m(Y)$$

$$+(p + q)m(A)\frac{\sum_{Y\cap A=\emptyset} m(Y)}{\sum_{Y\subseteq A} m(Y)} + (r\frac{\sum_{Y\cap A=A} m(Y)}{\sum_{Y\cap A\neq\emptyset} m(Y)} + s) \sum_{Y\cap A=\emptyset} m(Y).$$

[6] We have to note, that the coincidence of DRC with Dempster's rule, which is criticised and strictly rejected by Dezert and Smarandache is not considered in this comparison.

for $X \subset A$ whenever $Bel(A) > 0$ or $p = 0$ & $Pl(A) = 1$ or $p+q = 0$ & $Pl(A) > 0$ or $p+q+r = 0$ (i.e. $s = 1$); $m(X \wr A) = 0$ for $X \not\subseteq A$.

The above general formulas represent FCR for $p = 1$, BCR12 for $q = 1$, DRC for $r = 1$ and the plain BCR for $s = 1$. Admitting $p, q, r, s \in [0, 1]$ such that $p + q + r + s = 1$, we can consider the above general formulas as a definition of a new general BCR. Unfortunately such a general rule overtakes all negative properties of all four single rules from the least definition domain of BFR ($\{Bel | Bel(A) > 0\}$) to the greatest additional information. The rule is neither compatible with any of the classic combination rules.

As we have shown that DRC has all the above investigated properties better or equal to the properties of BFR and BCR12, we can improve the properties of the new general BCR by setting $p = q = 0$, hence we obtain simpler improved version of the general BCR for $r, s \in [0, 1]$ such that $r + s = 1$. The rule, of course, cumulate negative properties of DRC and of the plain BCR in general. In the special cases of $r = 1$ or $s = 1$, it coincides with DRC or with the plain BCR, respectively. We can correctly extend its definition domain using Desert-Smarandache idea $m(A|A) = 1$ whenever rule is not defined by the above formulas, i.e. $m(A \wr A) = 1$ when $Pl(A) = 0$ and $s \neq 1$ as it follows:

$$m(X \wr A) = (r + s) \sum_{Y \cap A = X} m(Y) + r \, \frac{\sum_{Y \cap A = X} m(Y)}{\sum_{Y \cap A \neq \emptyset} m(Y)} \sum_{Y \cap A = \emptyset} m(Y),$$

$$m(A \wr A) = (r + s) \sum_{Y \cap A = A} m(Y) + (r \, \frac{\sum_{Y \cap A = A} m(Y)}{\sum_{Y \cap A \neq \emptyset} m(Y)} + s) \sum_{Y \cap A = \emptyset} m(Y),$$

for $X \subset A$ whenever $Pl(A) > 0$ or $s = 1$;
$m(A \wr A) = 1$ when $Pl(A) = 0$ and $s \neq 1$; $m(X \wr A) = 0$ for $X \not\subseteq A$.

This general rule is defined for all classic BFs, it is weakly commutative in the sense of Section 6, its additional information is comparable with those of DRC. The rule preserves BBFs only for $r = 1$ (or in consequence with Bel_T [1, 2], in that case it preserves BBFs and coincides with DRC for BBFs for any $r, s \in [0, 1]$ such that $r + s = 1$). In general case the rule is not compatible with any classic rule of belief combination, for $r = 1$ it is compatible with Dempster's rule of combination and for $s = 1$ with Yager's and Dubois-Prade rules of combination.

9 Conclusion

The new belief conditioning rule — the plain BCR — was presented in this contribution. Properties of the plain BCR were compared with the properties of the classic BCRs including BCR12 — the most useful DSm BCR — applied to the classic belief functions. Dempster's rule of conditioning (DRC) has been shown to be better than the belief focusing rule (BFR) and BCR12 from the point of view of the investigated properties. Hence DRC and the plain BCR are recommended for belief conditioning, moreover DRC is compatible with with Dempster's rule of combination and the plain BCR is compatible with Yager's and Dubois-Prade rules of combination.

In the end a general BCR, which includes DRC and the plain BCR as its special cases, was defined and presented.

References

[1] Daniel M. (2005), Probabilistic Transformations of Belief Functions. In: Godo L. (Ed.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Proceedings ECSQARU 2005*, LNAI 3571, Springer-Verlag, 539–551.

[2] Daniel M. (2006) On Transformations of Belief Functions to Probabilities. *International Journal of Intelligent Systems* **21** No. 3, 261–282.

[3] Daniel M. (2007), The DSm Approach as a Special Case of the Dempster-Shafer Theory. In: K. Mellouli (Ed.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Proceedings ECSQARU 2007*, LNAI 4724, Springer-Verlag, 381–392.

[4] Daniel M. (2009), Analysis of DSm belief conditioning rules and extension of their applicability. Chapter 10 in: Smarandache F., Dezert J. (eds.), *Advances and Applications of DSmT for Information Fusion*, Volume 3. American Research Press (ARP), 323–343.

[5] Dezert J. (2002), Foundations for a New Theory of Plausible and Paradoxical Reasoning. *Information and Security, An International Journal* **9**.

[6] Dubois D., Prade H. (1988), Representation an combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, **4** 244–264.

[7] Jiroušek R., Vejnarová J. (1997), Uncertainty in Expert Systems. Chap. 2 in V. Mařík, O. Štěpánková, J. Lažanský, et al.: *Artificial Intelligence (2)*, (in Czech), Academia, Praha, 78–101.

[8] Shafer G. (1976), *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.

[9] Smarandache F., Dezert J. (2006), Belief Conditioning Rules, Chapter 9 in: Smarandache F., Dezert J. (eds.), *Advances and Applications of DSmT for Information Fusion*. Volume 2. American Research Press, Rehoboth, 237-268.

[10] Smets, Ph. (2005), Decision making in the TBM: the Necessity of the Pignistic Transformation. *Int. J. of Approximative Reasoning*, **38** 133–147.

[11] Yager R. R. (1987), On the Demspter-Shafer framework and new combination rules. *Information Sciences*, **41** 93–138.

[12] Yager R. R., Liu L. (eds) (2008), Classic works of the Dempster-Shafer theory of belief functions, Springer, Berlin, Heidelberg, 806 p.

The Role of Assumptions in Causal Discovery

Marek J. Druzdzel

Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA

Faculty of Computer Science, Białystok Technical University, Wiejska 45A, 15-351 Białystok, Poland

marek@sis.pitt.edu

Abstract

The paper looks at the conditional independence search approach to causal discovery, proposed by Spirtes *et al.* and Pearl and Verma, from the point of view of the *mechanism-based* view of causality in econometrics, explicated by Simon. As demonstrated by Simon, the problem of determining the causal structure from data is severely underconstrained and the perceived causal structure depends on the *a priori* assumptions that one is willing to make. I discuss the assumptions made in the independence search-based causal discovery and their identifying strength.

1 Introduction

An accepted scientific procedure for demonstrating causal relations is experimentation. If experimental manipulation of one variable (called the independent variable) results in a change in value of another variable (called the dependent variable), assuming an effective control for all possible intervening variables, one usually concludes that in the system under study the two variables stand in a causal relation with each other. Unfortunately, conducting such experiments is for many practical systems impossible, because of our inability to manipulate the system variables, forbidding costs of experimentation, or ethical considerations. Numerous examples of such systems are found in economics, medicine, meteorology, or social sciences. Still, one wants to predict the impact of policy decisions, such as whether to impose a tax, introduce or abolish the death penalty, or restrict smoking, on such variables as the gross national product, crime rates, or the number of lung cancer cases in the population. Where experimentation is impossible, one must rely on observations and assumptions in order to form a theory of causal interactions.

One discipline where much attention has been paid to model construction from observations is econometrics. Work in late 1940s and early 1950s (see for example [4] or [3]) concentrated on formulating economic theories in the form of systems of structural equations, i.e., equations describing mechanisms by which variables interact directly with each other. It was commonly believed that systems of structural equations should be formulated either entirely on the basis of

economic theory or economic theory combined with systematically collected statistical data for the relevant variables in the system. Construction of a system in the second case consisted of proposing a theoretical model, i.e., specifications of the form of the structural equations (including designation of the variables occurring in each of the equations) and then estimating the constant parameters from observations. The limits of such estimation raised the problem of "identifiability," i.e., whether it is theoretically possible, given prior knowledge about the functional forms of equations in a set of simultaneous equations, to determine unique values of parameters of these equations from observations. Simon [7] related the problem of identifiability to the causal structure of the system, showing theoretical conditions under which a structure is identifiable.

In their influential work, Spirtes *et al.* [9] and Pearl and Verma [5],[1] proposed that, under certain circumstances, observation is sufficient to determine all or part of the causal structure of a system. They have outlined methods for identifying the narrow class of causal structures (ideally a unique causal structure) that are compatible with particular observations. I will refer to the view of causality that underlies this work as *independence search-based* view of causality (or briefly ISC). As Simon [8] demonstrated, the problem of determining the causal structure from data is severely underconstrained and the perceived causal structure depends on the *a priori* assumptions that one is willing to make. From this point of view, there is little doubt that these new methods rest on some powerful identifying assumptions.

The goal of this paper is to explicate these assumptions, express them in terms of the earlier work in econometrics on structural equation models, and discuss their identifying strength. I will build on the results presented in [2], which reviews the mechanism-based view of causality (MBC) and shows a link between causal ordering and directed probabilistic graphs. The main conclusion resulting from this analysis is that with respect to the meaning of causality, the ISC and MBC views are almost identical. The power of the new methods rests on additional assumptions about causal relations that had not been made in econometrics. The two new powerful identifying assumptions are (1) that the causal structure is acyclic and (2) that each observed independence and dependence is a reflection of the causal structure and not merely coincidental (the latter called in the ISC view "faithfulness assumption"). With respect to the faithfulness assumption, the new, previously unexplored, element is dependence of causes conditional on a common effect.

The remainder of the paper is structured as follows. Section 2 starts with a brief review of the mechanism-based view of causality in directed probabilistic graphs. Section 3 offers a summary of the main assumptions made in the causal discovery work. Section 4 covers important concepts at the foundations of causal discovery: independence, conditioning, Markov condition, and faithfulness. It proposes a deterministic notion of independence and explains the link between this and the probabilistic view. Section 5 translates the assumptions in ISC into the MBC and explicates their identifying power.

[1] I will refer frequently to the book by Spirtes *et al.* [9] rather than to the work of Pearl and Verma, because I am more familiar with the former. I believe that for the purpose of this analysis, both approaches are equivalent. There are other, Bayesian approaches to causal discovery originating from the seminal work of Cooper and Herskovitz [1], which I will leave outside this discussion.

2 Mechanism-Based View Of Causality

The mechanism-based view of causality rests on the observation that individual causal mechanisms, while normally symmetric (e.g., forces are reciprocal), exhibit asymmetry when embedded in the context of a model. Simon [7] proposed a procedure for deriving a directed graph of interactions among individual variables, called *causal ordering*, and tied it to the econometric notion of structure. He postulated that when each of the equations in the model is structural and each of the exogenous variables is truly exogenous, the asymmetry reflects the causal structure of the system. Druzdzel and Simon [2] have shown the link between causal ordering and directed probabilistic graphs. I will briefly review the main results from that work.

The following theorem demonstrates that the joint probability distribution over n variables of a Bayesian network (BN) can be represented by a model involving n simultaneous equations with these n variables and n additional independently distributed latent variables.

Theorem 1 (representability) *Let $\mathcal{B}$ be a BN model with discrete random variables. There exists a simultaneous equation model $\mathcal{S}$, involving all variables in $\mathcal{B}$, equivalent to $\mathcal{B}$ with respect to the joint probability distributions over its variables.*

The following theorem establishes an important property of a structural equation model of a system with the assumption of causal acyclicity.

Theorem 2 (acyclicity) *The acyclicity assumption in a causal graph corresponding to a self-contained system of equations $\mathcal{S}$ is equivalent to the following condition on $\mathcal{S}$: Each equation $e_i \in \mathcal{S} : f(x_1, \ldots, x_n, \mathcal{E}_i) = 0$ forms a self-contained system of some order k and degree one, and determines the value of some argument x_j ($1 \leq j \leq n$) of f, while the remaining arguments of f are direct predecessors of x_j in causal ordering over $\mathcal{S}$.*

The last theorem binds causal ordering with the structure of a directed probabilistic graph.

Theorem 3 (causality in BNs) *A Bayesian belief network $\mathcal{B}$ reflects the causal structure of a system if and only if (1) each node of $\mathcal{B}$ and all its direct predecessors describe variables involved in a separate mechanism in the system, and (2) each node with no predecessors represents an exogenous variable.*

The above results show a link between structural equation models and causal graphs. They also make it clear that the former give a more general notion of structure than the letter. Directed probabilistic graphs are acyclic, while causal ordering in structural equation models can lead to cyclic structures. While equations can easily model dynamic processes with feedback loops, directed acyclic graphs can capture only their equilibrium states.

Theorem 3 demonstrates that directed arcs in BNs play a role that is similar in its representational power to the structure (presence or absence of variables in equations) of simultaneous equation models. The graphical structure of a BN, if given causal interpretation, is a qualitative specification of the mechanisms acting in a system.

3 Causal Discovery

Causal discovery in ISC is based on two axioms binding causality and probability. Informally, the first axiom, *causal Markov condition*, states that once we know all direct causes of an event, the event is probabilistically independent of its causal non-descendants. For example, suppose that we see a broken glass bottle on the bicycle path with small pieces of glass lying all around. Learning the cause of this broken bottle or that a piece from the bottle hurt a passing dog, does not change our expectation of a flat tire caused by the pieces of glass on the road.[2] The formal statement of the causal Markov condition is as follows:

> **Causal Markov Condition:** [9, page 54,] Let $\mathcal{G}$ be a causal graph with vertex set $\mathbf{V}$ and $\mathcal{P}$ be a probability distribution over the vertices in $\mathbf{V}$ generated by the causal structure represented by $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{P}$ satisfy the Causal Markov Condition if and only if for every W in $\mathbf{V}$, W is independent of $\mathbf{V}\backslash(\text{Descendants}(W) \cup \text{Parents}(W))$ given $\text{Parents}(W)$.

The second axiom, the *faithfulness condition*, assumes that all interdependencies observed in the data are structural, resulting from the structure of the causal graph, and not accidental (e.g., by some particular combination of parameter values that result in causal effects canceling out). Spirtes *et al.* demonstrate that purely accidental dependencies and independences have, under a wide class of natural distributions over the parameters, a probability of measure zero. The formal statement of the faithfulness condition is as follows:

> **Faithfulness Condition:** [9, page 56,] Let $\mathcal{G}$ be a causal graph and $\mathcal{P}$ a probability distribution generated by $\mathcal{G}$. $< \mathcal{G}, \mathcal{P} >$ satisfies the Faithfulness Condition if and only if every conditional independence relation true in $\mathcal{P}$ is entailed by the Causal Markov Condition applied to $\mathcal{G}$.

One of the consequences of the causal Markov condition in combination with the faithfulness condition is conditional dependence: all causal predecessors of an observed variable v become probabilistically dependent conditional on v. Suppose that while riding a bicycle we get a flat tire. This makes all possible causes of the flat tire probabilistically dependent conditional on the flat tire. Observing pieces of glass on the road, for example, makes thorns less likely (the glass "explains away" the thorns).

Markov and faithfulness conditions bind causality with probability and along with other assumptions, such as acyclicity of the causal structure, reliability of the statistical tests applied, or independence of error terms, place constraints on the causal structure. The constraints provide clues to the causal structure that generated the observed patterns of interdependencies. Spirtes *et al.* show that given their assumptions, they are often able to reconstruct from a set of

[2] Many of these properties of causes have been long known. Reichenbach described "causal forks" consisting of a cause and two or more effects. The effects are normally probabilistically dependent because of the common cause, but this dependence vanishes if we condition on the cause [6, page 158,]. The causal Markov condition is not completely uncontroversial. Salmon [6] postulates the existence of "interactive forks," that violate the causal Markov condition. Spirtes *et al.* give an appealing explanation of Salmon's examples and postulate that interactive forks do not exist, at least in the macroscopic world [9, Section 3.5.1,].

observations a unique causal structure of the system that generated them. The search for that causal structure is a search for the class of faithful models that are structurally able to generate the observed independences, and sometimes this search provides a unique structure.

4 Independence, Conditioning, Markov Condition, and Faithfulness

This section builds a bridge from the MBC to the ISC view of causality by introducing a deterministic notion of independence between a system's variables. This is a purely theoretical exercise that allows to talk about dependences among variables in a system of simultaneous structural equations. Please note that the concept of causal ordering, as explicated by Simon, operates on systems of simultaneous structural equations with no notion of uncertainty. Uncertainty enters these systems through variability of exogenous variables (error terms are simply exogeous variables on the par with other exogenous variables).

4.1 Deterministic Independence

I propose to base the deterministic definition of independence on the notion of dimensionality of the Cartesian product of variables. Followings the conventions in physics and mathematics, I define the dimension of a space roughly as the minimum number of coordinates needed to specify every point within it. A Cartesian product of n independent variables has dimensionality n, for, as each of the variables can vary independently over its domain, the points in this product cover an n dimensional space. If there is any interdependency among the variables, there will be loss in the dimensionality of this space. For example, if the element binding the two variables is an equation describing a unit circle, all we need to specify a point in this space is the polar coordinate angle. The Cartesian product of two independent variables forms a plane. If these two variables are dependent, then the domain of their Cartesian product will have a lower dimensionality and will be a line. The value of one of the variables puts a constraint on the value of the other.

Definition 1 (independence) *Sets of variables $\mathcal{X}$ and $\mathcal{Y}$ in a simultaneous equation model $\mathcal{S}$ are independent if the dimensionality of the Cartesian product of the variables in $\mathcal{X} \cup \mathcal{Y}$ is equal to the sum of dimensionalities of Cartesian products of variables in $\mathcal{X}$ and $\mathcal{Y}$ separately.*

Loss of dimensions is caused by functional relations that bind variables between the sets $\mathcal{X}$ and $\mathcal{Y}$. Each functional relation causes, in general, loss of one dimension. Because the exercise is theoretical, I will leave out of this paper the question how to test for deterministic independence in practice, along the lines of testing probabilistic independence.

4.2 Conditioning

Conditioning within a system of simultaneous equations means selecting a subset of observations that fulfills some specified condition. Such a condition forms

a constraint on the values that a measured variable or a set of measured variables can take in the selected subset. Typically, one requires the value of a variable to be equal to some constant value. Conditioning is a passive way of "experimenting" with the system without modifying its causal structure. One selects those instances of the system's output that produce a specified value. If we condition on, for example, $x_i = x_{i_0}$, then we add to the system an additional constraint

$$x_i = x_{i_0} \ . \tag{1}$$

It is important to distinguish conditioning from direct manipulation of x_i, which is referred to in econometrics by *change in structure*. A change in structure is represented by replacing the equation that is made inactive by an equation describing the manipulation. In this case, one would replace the equation e_i that determines the value of x_i by the equation $x_i = x_{i_0}$. In conditioning, on the other hand, the selected data set needs to satisfy the equation e_i and, in addition, Equation 1. Conditioning on one variable reduces, thus, the system from a self-contained set of n equations with n variables to a set of $n+1$ equations with n variables (or, if we choose to replace x_i by a constant, n equations with $n-1$ variables), a system that is overconstrained. (This system still has solutions — these are the observed data points.)

4.3 Markov Condition

It turns out that in deterministic models, Markov condition can be derived rather than assumed and the following theorem can be proven.

Theorem 4 (Markov condition) *Let $\mathcal{S}$ be a simultaneous equation model with n variables $\mathbf{V}$ and n independent error variables. Let $\mathcal{G}$ be a directed acyclic graph with vertex set $\mathbf{V}$ reflecting causal ordering over variables $\mathbf{V}$ in $\mathcal{S}$. For every $w \in \mathbf{V}$, w is independent of $\mathbf{Z} \equiv \mathbf{V} \backslash$(Descendants($w$) $\cup$ Parents(w)) given Parents(w).*

The theorem shows that the Markov condition is a simple consequence of the fact that the system is modeled by a set of simultaneous structural equations.

Theorem 2 shows that under the assumption of acyclicity, each of the equations determines one variable. Let equation e_i determine the variable x_i and precede (in the causal ordering over the model) all equations e_j, such that $i < j$. Because none of the equations e_k, such that $k < i$, contained x_i, each remains unchanged when we condition on x_i. Each of those equations e_k such that $i < k$ that contained x_i will now contain one fewer variable. This will lead to making the causal path from the predecessors of x_i to its causal successors inactive: note that as x_i becomes constant, none of the equations for the causal successors of x_i will depend on causal predecessors of x_i through x_i (they may, of course, depend through other paths).

4.4 The Faithfulness Assumption

I propose the following deterministic definition of faithfulness.

Definition 2 (faithfulness) *A structural equation model $\mathcal{S}$ is* faithful *with respect to its structure if and only if every independence between sets of variables*

in $\mathcal{S}$ is entailed by the structure of $\mathcal{S}$ (i.e., by the presence and the absence of variables in individual equations in $\mathcal{S}$).

What this definition requires practically is that the model not contain equations that structurally look as if they were putting a constraint on a variable or a set of variables, but where in reality, the actual functional form and the actual values of the coefficients imply no constraint. Unfaithfulness may happen when a variable is present in an equation, but the coefficient of that variable is zero or becomes zero when influence through different paths is being computed (i.e., when the total effect of a variable on another variable through different paths "cancels out").

There are dependencies that do not result in loss of dimensionality, such as Peano or Sierpiński curves, or even the simple absolute value function. However, one has to remember that there are dependences that do not result in probabilistic dependence, for example deterministic dependences, excluded by the faithfulness axiom in the ISC approach.

4.5 Useful Properties of Causal Graphs

I report three properties of the relation between causal ordering and independence. Proofs are quite straightforward and omitted due to space constraints.

Theorem 5 (causal dependence) *If y precedes x in causal ordering, then y and x are dependent.*

Theorem 6 (spurious dependence) *If z precedes both x and y in the causal ordering, then x and y are dependent.*

One of most useful conclusions that can be drawn from conditioning is conditional dependence. Conditioning on a variable in a simultaneous equation model yields a data set in which all variables that are causal predecessors of that variable are dependent, contrary to the situation before conditioning, where exogenous variables in the system under study were independent by assumption. This observation shows that conditioning on a set of variables allows one to draw inferences about the causal ordering of variables, namely to discriminate, under certain circumstances, between causal predecessors and causal successors of the variables that were conditioned on. This property is captured by the following theorem.

Theorem 7 (conditional dependence) *Let $\mathcal{S}$ be a self-contained simultaneous equation model. Let Ψ be the set of causal predecessors of a variable x. Given the faithfulness assumption, any two subsets of variables $\mathcal{Y}, \mathcal{Z} \in \Psi$, are dependent conditional on x.*

The above three theorems show that causal ordering and interdependence are related. Causal ordering of the variables in a system of equations will result in a pattern of interdependencies in the observed data. This pattern, in turn, will give clues to the causal ordering or, more exactly, to the structural equations of the system.

5 Assumptions in Causal Discovery

Using elementary algebraic considerations, Simon [8] demonstrated that the problem of determining a causal structure, either from experimental or observational data, is severely underconstrained. The way one perceives the causal structure of a system is strongly dependent on the assumptions that one is willing to make. In particular, one might assume that a causes b only from an observed correlation between a and b, if one is willing to make the assumptions of time precedence and causal sufficiency (the latter excludes the possibility of a common cause) [8]. Similarly, one may be reluctant to accept even an experimental demonstration of causation, if one rejects critical assumptions about the experimental setup. It is, therefore, essential to state explicitly the assumptions made and provide the motivation for their validity.

In this section, I will outline the identifying information supplied by each of the assumptions made in causal discovery and the exact gains for discovering causality. I will use the structure matrix notation introduced in [2] and reproduced below to show the gains from each of the assumptions in terms of the number of coefficients of structural equations that are determined by the assumption. Some of the assumptions work in combination, and it is, therefore, difficult to assess the net gain obtained by each separately.

5.1 Initial Observations

The following definition, reproduced from [2], introduces a convenient notation for the structure of equations in simultaneous equation models.

Definition 3 (structure matrix) *The structure matrix A of a system $\mathcal{S}$ of n simultaneous structural equations e_1, e_2, ..., e_n with n variables x_1, x_2, ..., x_n is a square $n \times n$ matrix in which element a_{ij} (row i, column j) is non-zero if and only if variable x_j participates in equation e_i. Non-zero elements of A will be denoted by X and zero elements by* 0.

Our starting point is the observation that any system can be modeled by n measured variables (x_1, x_2, ..., x_n) and n unmeasured latent variables ($\mathcal{E}_1$, $\mathcal{E}_2$, ..., $\mathcal{E}_n$), called error terms (note that I am not making any assumptions about their interdependence). If we denote the ith measured variable by x_i and the ith error term by $\mathcal{E}_i$, we can write the following structure matrix A for the set of $2n$ simultaneous structural equations with $2n$ variables. A solution of this set for any given set of values of the exogenous variables $\mathcal{E}_1$, $\mathcal{E}_2$, ..., $\mathcal{E}_n$ describes a single observed data point.

$$\begin{bmatrix}
 & x_1 & \dots & x_n & \mathcal{E}_1 & \dots & \mathcal{E}_n \\
(e_1) & a_{11} & \dots & a_{1n} & a_{1n+1} & \dots & a_{12n} \\
(e_2) & a_{21} & \dots & a_{2n} & a_{2n+1} & \dots & a_{22n} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
(e_n) & a_{n1} & \dots & a_{nn} & a_{nn+1} & \dots & a_{n2n} \\
(e_{n+1}) & a_{n+11} & \dots & a_{n+1n} & a_{n+1n+1} & \dots & a_{n+12n} \\
(e_{n+2}) & a_{n+21} & \dots & a_{n+2n} & a_{n+2n+1} & \dots & a_{n+22n} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
(e_{2n}) & a_{2n1} & \dots & a_{2nn} & a_{2nn+1} & \dots & a_{2n2n}
\end{bmatrix} \tag{2}$$

5.2 Acyclicity of the Causal Structure

The acyclicity assumption is probably the strongest assumption made in ISC causal discovery. It technically amounts to assuming that there are no feedback

loops in the causal graph of the system. The implication of this assumption for a simultaneous structural equation model has been captured by Theorem 2. Every equation in such a model determines the value of exactly one endogenous variable.

Before showing the implications of the acyclicity assumption for causal discovery, I will rearrange the coefficients of the structure matrix to a form convenient in causal discovery. Without loss of generality we are free to assume that row i $(i = 1, \ldots, n)$ in (2) represents equation e_i that determines the value of variable x_i, and row $n + j$ $(j = 1, \ldots, n)$ represents equation e_{n+j} that determines the value of the error variable $\mathcal{E}_j$. Also, column i $(i = 1, \ldots, n)$ of the matrix will contain coefficients for the variable x_i and column $n+j$ $(j = 1, \ldots, n)$ will contain coefficients of the error variable $\mathcal{E}_j$. Mathematically this assumption amounts to rearranging the structure matrix by row and column exchanges (renaming the variables and the equations), which, as shown in [7], preserves the causal structure of the system.

The following three properties hold in this rearranged matrix A:

Property 1 (diagonal elements) $\forall_{1 \leq i \leq 2n}\ a_{ii} \neq 0$

Because I assumed that equation e_i determines variable x_i, and, therefore, x_i must be present in e_i, each diagonal element of A must be non-zero (i.e., $\forall_{1 \leq i \leq n}\ a_{ii} \neq 0$). The same holds for the error variables $\mathcal{E}_i$.

Property 2 (off-diagonal elements) *There are at least* $2n(2n-1)/2$ *zeros among off-diagonal elements of A. All non-zero off-diagonal elements in A represent direct causal predecessors of the diagonal element of the same row.*

In the proof of Theorem 2 [2], I demonstrate that another implication of the acyclicity assumption is that the structure matrix is triangular and, therefore, contains at least $2n(2n-1)/2$ zeros. The location of these zeros is only partly disclosed and can be retrieved only in combination with the properties of the observed data and other assumptions during the discovery process.

By Theorem 2, all variables that participate in an equation, except the one that is determined by the equation, are direct causal predecessors of that variable. By Property 1, the diagonal elements denote the variables that are being determined, therefore, it follows that all non-zero off-diagonal elements represent direct causal predecessors of the diagonal elements. Note that no assumptions have been made so far about interdependence of error variables and each of the equations e_{n+i} $(i = 1, \ldots, n)$ can model dependencies among these.

Property 3 (acyclicity) $\forall_{i \neq j}\ a_{ij} \neq 0 \Longrightarrow a_{ji} = 0$

$a_{ij} \neq 0$ implies that x_j is a direct predecessor of x_i and $a_{ji} \neq 0$ would imply that x_i is a direct predecessor of x_j, which then implies a cycle in the causal graph. Note that Property 3 captures only cycles of degree two. It is possible to capture cycles of higher degrees, although the conditions for these become increasingly complex.

5.3 Causal Sufficiency

The assumption of causal sufficiency[3] is equivalent to the assumption of independence of exogenous variables $\mathcal{E}_i$. Independence of exogenous variables amounts to assuming that half of the $2n$ equations contain just one variable, namely one of the n error terms. As the remaining n equations each involves exactly one distinct variable of the n error variables, we get $3n^2 - 2n$ structural zeros.

$$\begin{bmatrix}
 & x_1 & x_2 & \dots & x_n & \mathcal{E}_1 & \mathcal{E}_2 & \dots & \mathcal{E}_n \\
(e_1) & X & 0 & \dots & 0 & X & 0 & \dots & 0 \\
(e_2) & a_{21} & X & \dots & 0 & 0 & X & \dots & 0 \\
(e_3) & a_{31} & a_{32} & \dots & 0 & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
(e_n) & a_{n1} & a_{n2} & \dots & X & 0 & 0 & \dots & X \\
(e_{n+1}) & 0 & 0 & \dots & 0 & X & 0 & \dots & 0 \\
(e_{n+2}) & 0 & 0 & \dots & 0 & 0 & X & \dots & 0 \\
(e_{n+3}) & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
(e_{2n}) & 0 & 0 & \dots & 0 & 0 & 0 & \dots & X
\end{bmatrix}$$

For the sake of simplicity of the subsequent discussion, we can remove the error term parts of the above structure matrix (note that the removed parts contain no unknown values of parameters), obtaining:

$$\begin{bmatrix}
 & x_1 & x_2 & x_3 & \dots & x_n \\
(e_1) & X & 0 & 0 & \dots & 0 \\
(e_2) & a_{21} & X & 0 & \dots & 0 \\
(e_3) & a_{31} & a_{32} & X & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots \\
(e_n) & a_{n1} & a_{n2} & a_{n3} & \dots & X
\end{bmatrix} \tag{3}$$

In the structure matrix above, I have assumed that all zeros are located above the diagonal to show graphically the number of structural zeros obtained by the acyclicity assumption. In fact, the location of zeros is not disclosed a-priori and is only constrained by Property 3. The actual inference from the observed pattern of interdependencies concentrates on determining for each of the remaining $n(n-1)/2$ coefficients in (3) whether it is zero or non-zero.

6 Conclusion

Because the problem of causal inference from observations is severely underconstrained, the perceived causal structure depends on the *a priori* assumptions that one is willing to make. This paper has explicated the assumptions made in the causal discovery work (ISC view) and expressed them in terms of the earlier work in econometrics on structural equation models (mechanism-based view). I discussed the identifying strength of each of the assumptions in terms of the number of structural zeros and non-zeros that can be implied in the structure matrix.

The power of the ISC methods seems to rest on additional assumptions about causal relations that had not been made in econometrics. The two new powerful

[3] This assumption can be relaxed in ISC causal discovery — some search algorithms proposed by Spirtes *et al.* allow for discovery of models that are not causally sufficient. In this case, the algorithm suggests possible common causal predecessors of any pair of the measured variables [9, Chapter 6,].

identifying assumptions are acyclicity of the causal structure and the assumption that each observed independence and dependence is a reflection of the causal structure of the system and is not merely coincidental (the latter called by Spirtes *et al.* "faithfulness assumption"). With respect to the faithfulness assumption, the new, previously unexplored, element is dependence of causes conditional on a common effect: a conditional dependence observed in this case is assumed to be structural and allows for distinguishing between predecessors and successors of the node conditioned on.

Acknowledgments

I am indebted to the late Herb Simon for his guidance and, especially, his contributions to my understanding of the mechanism-based view of causality. Had he been alive, he would have been a co-author on this paper. I thank Peter Spirtes, Richard Scheines, Clark Glymour, Chris Meek and Greg Cooper for inspiring discussions on the subject of causal discovery and helpful comments on this work.

References

[1] Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[2] Marek J. Druzdzel and Herbert A. Simon. Causality in Bayesian belief networks. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, pages 3–11, San Francisco, CA, 1993. Morgan Kaufmann Publishers, Inc.

[3] William C. Hood and Tjalling C. Koopmans, editors. *Studies in Econometric Method. Cowles Commission for Research in Economics. Monograph No. 14.* John Wiley & Sons, Inc., New York, NY, 1953.

[4] Tjalling C. Koopmans, editor. *Statistical Inference in Dynamic Economic Models. Cowles Commission for Research in Economics. Monograph No. 10.* John Wiley & Sons, Inc., New York, NY, 1950.

[5] Judea Pearl and Thomas S. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *KR–91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, Cambridge, MA, 1991. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

[6] Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World.* Princeton University Press, Princeton, NJ, 1984.

[7] Herbert A. Simon. Causal ordering and identifiability. In Hood and Koopmans [3], chapter III, pages 49–74.

[8] Herbert A. Simon. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 49(267):467–479, September 1954.

[9] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer Verlag, New York, 1993.

Bridges Between Contextual Linguistic Models of Vagueness and T-norm Based Fuzzy Logic

Christian G. Fermüller
Theory and Logic Group
Vienna University of Technology
chrisf@logic.at

Christoph Roschger
Theory and Logic Group
Vienna University of Technology
roschger@logic.at

Abstract

Linguistic models of vagueness usually record contexts of possible precisifications. A link between such models and fuzzy logic is established by extracting fuzzy sets from context based word meanings and analyzing standard logical connectives in this setting. In a further step Lawry's voting semantics for fuzzy logics is used to re-interpret standard t-norm based truth functions from the point of view of context update semantics.

1 Introduction

Vagueness is a significant and ubiquitous phenomenon of human communication. Adequate models of reasoning with vague information are not only of perennial interest to philosophers and logicians (see, e.g., [15, 14, 31, 5, 28] and references there), but are also a topic of current linguistic research. Of particular interest from a logical point of view are approaches to formal semantics of natural language that can be traced back to Richard Montague's ground breaking work, firmly connecting formal logic and linguistics (see, e.g., [24, 12]).

At a first glimpse it seems that most linguistic models of vagueness are *incompatible* with the degree based approach offered by fuzzy logic. In particular, there are indeed good reasons why we should not simply replace Montague's type $\mathbf{t} = \{0, 1\}$ for sentences, i.e. the classical truth values *false* and *true*, by the unit interval $[0, 1]$ if we aim at a realistic and adequate model of meaning in natural language. What is rather needed, as is made clear e.g. in [25, 3, 1, 16, 18], are models that systematically take into account contexts of utterance that record relevant possible precisifications of vague word meanings. Our aim is to bridge the seemingly wide gap between such linguistic models and fuzzy logic by demonstrating how *fuzzy sets* can be systematically extracted from the meaning of predicates in a given context. To make this concrete we will refer to a specific linguistic framework—dynamic context semantics—as used by Chris Barker [1] for the analysis of vagueness. Building on this connection between contexts and fuzzy sets we will also investigate how the truth functional approach of fuzzy logic can be justified under certain conditions. Again, we will refer to

a specific example, namely Lawry's [19] voting semantics, to illustrate how a corresponding re-interpretation of logical operators could look like.

2 Linguistic approaches to vagueness

Linguists, like logicians, often focus on *predicates* and *predicate modifiers* in modeling the semantics of vague language. It is impossible to provide a survey on the relevant literature that does justice to all linguistic approaches to vagueness in short space.[1] For our purpose it suffices to note that there seems to be wide agreement that adequate truth conditions for vague sentences have to refer not only to fixed lexical entries, but also to *contexts of utterance* that may be identified with sets of contextually relevant possible *precisifications*. Indeed, many authors take it for granted that a realistic and complete formal semantics of natural languages has to take into account the context dependence of truth conditions, anyway, e.g., to be able to resolve ambiguities and to handle anaphora. However, some care has to be taken, since 'context' can mean different things here that may operate on different levels. For example, it is obviously relevant to know, whether in applying the adjective *tall* the reference is to trees in a forest, to basket ball players, to women in central Europe, to school kids, or to *a tall story*. But even if, say, it is clear that the *general context* of asserting *Jana is tall* is a discussion about my students and not about basket players, arguably something like Lewis's *conversational score* [20] (cf. also [28]) is needed in addition to understand whether *Jana is tall* is meant to communicate information about Jana's height to someone who doesn't know her or whether speaker and hearer both have precise common information about Jana's height and the speaker intends to establish a standard of tallness by making this utterance. Reference to such 'conversational contexts of possible precisifications' is convincingly argued to be an essential ingredient of adequate models of communication with vague notions and propositions (see, e.g., [25, 3, 1, 16, 28]).

Instead of surveying the mentioned arguments, we will illustrate the versatile use of contexts in formal semantics by outlining just one particular, rather recent approach, due to Chris Barker [1]. This will serve as motivation and bridgehead—to stick with the metaphor in the title of this contribution—for exploring connections to fuzzy logic in the following sections. Barker casts his analysis of various linguistic features of vagueness in terms of so-called *dynamic semantics* (see [10]), that has been successfully employed to handle, e.g., anaphora. In this approach the meaning $[\![\phi]\!]$ of a declarative sentence (propositional expression) ϕ is given by an *update function* operating on the set of contexts. As already indicated above, semantic theories differ in their intended meaning and formal manifestation of the notion of contexts. Barker [1], following Stalnaker [29], identifies a context with a set of 'worlds', where in each world the extension of all relevant predicates with respect to the actual universe of discourse is *completely precisified*; i.e., each (relevant) atomic proposition is either true or false in a given world. For gradable adjectives these precisifications are specified by a *delineation* δ that, for each world, maps every gradable adjective—or more precisely: every reference to the meaning of a gradable adjective—into a particular value or degree of a corresponding scale. These values represent

[1] For this we refer to the handbook article [26], but also to the classic monograph [25], the more recent papers [1, 16, 18] and the references there.

local standards of acceptance. For instance, if $\delta(c)$ is the delineation function associated with world c, then $d = \delta(c)(^{\uparrow}[\![tall]\!])$ yields the standard of tallness in c expressed, say, in cm; i.e. every individual that is at least d cm tall in c will be accepted as *tall* in c.

In fact, only a simple form of update functions is needed; namely *filters*, where $[\![\phi]\!](C) \subseteq C$ holds for all contexts C—the result $[\![\phi]\!](C)$ being the set of worlds in C that survive the update of C with the assertion that ϕ. This observation entails that dynamic semantics is just a notational variant of a more traditional specification of 'truth at a world': ϕ is true (accepted) at c if $[\![\phi]\!](\{c\}) = \{c\}$ and ϕ is false (rejected) at c if $[\![\phi]\!](\{c\}) = \{\}$. Moreover, we assume that every world c of a given context C refers to the *same domain* (relevant universe of discourse) D_C.

Gradable predicates, like *tall*, express a relation involving degrees and individuals. The denotation of *tall* is modeled by a function tall such that $\mathsf{tall}(d, \mathsf{a})$ returns the set of worlds in which the individual a is at least d cm tall. Accordingly Barker presents the (dynamic) *meaning* of *tall* by[2]

$$[\![tall]\!] \quad =_{df} \quad \lambda x \lambda C.\{c \in C : c \in \mathsf{tall}(\delta(c)(^{\uparrow}[\![tall]\!]), x)\}$$

Among other features, this semantic setup allows Barker to capture the intuitive difference in the meaning of the modifiers *very*, *definitely*, and *clearly*. To define $[\![very]\!]$ an underlying relation very over degrees is used, such that $\mathsf{very}(s, d, d')$ holds iff the difference between d and d' is larger than the (vague, i.e., world dependent) standard s:

$$\begin{aligned}[\![very]\!] \quad =_{df} \quad \lambda\alpha\lambda x\lambda C.\{c \in \alpha(x)(C) : \exists d(c[d/\alpha] \in \alpha(x)(C) \wedge \\ \mathsf{very}(\delta(c)(^{\uparrow}[\![very]\!]), \delta(c)(^{\uparrow}\alpha), d)\}\end{aligned}$$

where $c[d/\alpha]$ denotes a world that is like c, except for setting $\delta(c)(^{\uparrow}\alpha) = d$. E.g., in $c[185cm/[\![tall]\!]]$ the standard of tallness is $185cm$. Thus $[\![Ann\ is\ very\ tall]\!] = ([\![very]\!]([\![tall]\!]))(\mathsf{Ann})$ is a filter (update) that is survived by exactly those worlds of a given context where *Ann* exceeds the standard of tallness by at least some amount s. This amount s not only depends on the meaning of *tall* and *very*, but also on the world itself. Thus the vagueness of *very* is modeled by a twofold context dependence: the meaning of *very* may obviously vary from context to context, but even within a fixed context different worlds may have different standards of accepting that an individual is *very tall*, granted that it is *tall*.

Note that, on the level of an individual world c, the update function for *very* refers only to information pertaining to c. In contrast, Barker suggests to model *definitely* as a type of modal operator:

$$[\![definitely]\!] =_{df} \lambda\alpha\lambda x\lambda C.\{c \in \alpha(x)(C) : \forall d(c[d/\alpha] \in C \rightarrow c[d/\alpha] \in \alpha(x)(C))\}$$

This means that a world $c \in C$ survives the update with $[\![Ann\ is\ definitely\ tall]\!]$ iff all worlds in C in which Ann has the same height as in c judge Ann as tall according to their local standard.[3]

[2] In fact Barker does not distinguish between $[\![tall]\!]$ and the purely referential use of it. Our notation is meant to indicate that the circularity is of a harmless type.

[3] Note that there might be uncertainty about Ann's height. I.e., Ann may have different heights in different worlds. Therefore *definitely tall* is not just equivalent to '*tall* in all worlds of the context'.

Finally, essential elements of $[\![very]\!]$ and $[\![definitely]\!]$ are combined in the following suggestion for the meaning of *clearly*:[4]

$$[\![clearly]\!] \quad =_{df} \quad \lambda\alpha\lambda x\lambda C.\{c \in \alpha(x)(C) : \mathsf{very}(\delta(c)(^{\uparrow}[\![clearly]\!]), \max_{\alpha}, \max_{C})\}$$

where $\max_{\alpha} = \{d : c[d/\alpha] \in \alpha(x)\}$ and $\max_{C} = \{d : c[d/\alpha] \in C\}$. The reference to $[\![clearly]\!]$ in the first argument of the relation very entails that, while the same comparison relation is used, the (world dependent) amount that the difference between the second and the third value has to exceed, may be different for *clearly* and *very*, respectively. However the essential difference between $[\![very]\!]$ and $[\![clearly]\!]$ is another one: while for *very tall* one compares the local standard of tallness with the local value for an individuals' height in each world, *clearly tall* involves a comparison of the highest standard of tallness in the whole context with the maximal height that the individual may have according to any world of the context.

3 Extracting fuzzy sets from contexts

Our main pillar in building a bridge between linguistics and fuzzy logics consists in connecting the meaning of predicates like *tall* with fuzzy sets. We define logical operators *and*, *or*, and *not* directly on predicates[5] in a natural way and explore how they relate to the corresponding operations on fuzzy sets. Note that linguists may seek to preserve the difference between statements like *Jana is tall and clever* and *Jana is tall and Jana is clever*, respectively. However, it will be straightforward to lift our analysis of predicate operators to the propositional level.

We introduce the notion of an *element filter*. These are filters parameterized by a domain element. Element filters that we have already encountered are e.g. $[\![tall]\!]$ but also $[\![very]\!]([\![tall]\!])$, where for a domain element x both $[\![tall]\!](x)$ and $([\![very]\!]([\![tall]\!]))(x)$ are filters.

Given a context C we can extract a fuzzy set from the meaning $\alpha = [\![A]\!]$ of a predicate A by applying for each domain element x the filter $\alpha(x)$ to C and measuring the amount of surviving worlds of C. For simplicity we stipulate contexts to be finite sets of worlds and identify fuzzy sets with their membership functions to obtain the following:

Definition 1. *Let C be a context with domain D_C and α an element filter. Then the fuzzy set $[\alpha]_C$ is given by*

$$[\alpha]_C : D_C \to [0,1] : \quad x \mapsto \frac{|\alpha(x)(C)|}{|C|}$$

Note that the collection of fuzzy sets $[\alpha]_C$ for all relevant element filters α carries less information that C itself. This will get apparent when we compare logical operators defined on predicates with corresponding operations on fuzzy sets.

Extending the framework of Barker, we model compound predicates (like *tall and clever*), built up from logically simpler predicates (*tall*, *clever*), in a straightforward manner:

[4] Our version of $[\![clearly]\!]$ differs in inessential details from Barker's in [1].

[5] For brevity we focus on monadic predicates, but the concepts can easily be extended to relations of higher arity.

Definition 2.

- $[\![and]\!] =_{df} \lambda\alpha\lambda\beta\lambda x\lambda C.\alpha(x)(C) \cap \beta(x)(C)$
- $[\![or]\!] =_{df} \lambda\alpha\lambda\beta\lambda x\lambda C.\alpha(x)(C) \cup \beta(x)(C)$[6]
- $[\![not]\!] =_{df} \lambda\alpha\lambda\beta\lambda x\lambda C.C\backslash(\alpha(x)(C))$

Note that in the above definition $\alpha = [\![A]\!]$ and $\beta = [\![B]\!]$ are element filters representing the meaning of the predicates A and B, respectively. Using the usual infix notation, $[\![A\ and\ B]\!]$ is an element filter as well. In general, applying $[\![A\ and\ B]\!]$ is not equivalent to applying the element filters $[\![A]\!]$ and $[\![B]\!]$ consecutively. We may additionally define

- $[\![and^*]\!]$[7] $=_{df} \lambda\alpha\lambda\beta\lambda x\lambda C.[\![B]\!](x)([\![A]\!](x)(C))$

Then $[\![A\ and^*\ B]\!]$ is, in general, different from $[\![A\ and\ B]\!]$ (and from $[\![B\ and^*\ A]\!]$).

The membership degree of x in the fuzzy set $[A\ and\ B]_C$[8] is determined by applying the filter $[\![A\ and\ B]\!](x)$ to the context C and calculating the fraction of worlds in C that survive this update. Proceeding a step further on our bridge from linguistics to fuzzy logics, the question arises if we can determine $[A\ and\ B]_C(x)$ from the membership degrees $[A]_C(x)$ and $[B]_C(x)$ alone. This, of course, would give us a fully truth-functional semantics for *and*, *or*, and *not*. However, fuzzy sets abstract away from the internal structure of contexts that may show various possible dependencies of worlds. We illustrate this by the following example.

Let C be a context consisting of the five possible worlds c_1 to c_5 as in Table 1. Furthermore, let $[\![jana]\!] = \mathsf{j}$ be a domain element and let tall, clever, and heavy be the denotations of the unary predicates *tall*, *clever*, and *heavy*, respectively, just as already demonstrated for tall and *tall* in Section 2.

c	$\delta(c)(\uparrow[\![tall]\!])$	$\mathrm{maxd}^{\mathsf{j}}_{\uparrow[\![tall]\!]}$	$\delta(c)(\uparrow[\![clever]\!])$	$\mathrm{maxd}^{\mathsf{j}}_{\uparrow[\![clever]\!]}$	$\delta(c)(\uparrow[\![heavy]\!])$	$\mathrm{maxd}^{\mathsf{j}}_{\uparrow[\![heavy]\!]}$
c_1	170	175	100	105	80	75
c_2	160	170	120	125	75	70
c_3	170	180	100	95	90	100
c_4	180	175	105	100	85	75
c_5	170	165	110	115	70	65

with maxd^x_p denoting the maximum degree to which to individual x fulfills the predicate referenced by p.

Table 1: Example Context C

Then $[\![heavy]\!]$ is an element filter with $[\![heavy]\!](\mathsf{j})(C) = \{c_3\}$. Accordingly, $[heavy]_C(\mathsf{j}) = 1/5$. Likewise we have $[clever]_C(\mathsf{j}) = [tall]_C(\mathsf{j}) = 3/5$. Since these latter are equal, also the membership degrees of j in the fuzzy sets $[tall\ and heavy]_C$ and $[clever\ and heavy]_C$. respectively, had to be equal if the (context update) meaning of *and* were truth functional. But $[\![tall\ and heavy]\!](\mathsf{j})(C) = \{c_3\}$, thus $[tall\ and heavy]_C(\mathsf{j}) = 1/5$, while $[clever\ and heavy]_C(\mathsf{j}) = 0$. As we see, by extracting the three fuzzy sets from the corresponding element filters we lose the

[6] In natural language one can also find *exclusive* disjunction, e.g. *Jana is either tall or clever (but not both).* He we focus on *inclusive* disjunction as this directly corresponds to disjunction as it is normally used in logics.

[7] Arguably, *and** corresponds to certain uses of *and even* and of *but*, respectively.

[8] For the sake of readability we write $[X]_C$ instead of $[[\![X]\!]]_C$.

information about the specific overlap of the corresponding updates in the given context.

The following bounds encode our best knowledge about membership degrees for fuzzy sets extracted from to composite predicates with respect to membership degrees referring to the corresponding components.

Theorem 1. *Let C be a context, $d \in D_C$, and let $\alpha = [\![A]\!]$ and $\beta = [\![B]\!]$ be two element filters. Then the following bounds are tight:*

- $\max\{0, [\alpha]_C(d) + [\beta]_C(d) - 1\} \leq [A \,and\, B]_C(d) \leq \min\{[\alpha]_C(d), [\beta]_C(d)\}$
- $\max\{[\alpha]_C(d), [\beta]_C(d)\} \leq [A \,or\, B]_C(d) \leq \min\{1, [\alpha]_C(d) + [\beta]_C(d)\}$
- $[not\, A]_C(d) = 1 - [\alpha]_C(d)$

Proof. The value $1 - [\alpha]_C(d)$ for negation follows directly from the relevant definitions.

For conjunction and disjunction we focus on the extremal cases: the sets $\alpha(d)(C)$ and $\beta(d)(C)$ may either be 'as disjoint as possible' or one set may contain the other one. In the latter case we have $\min\{[\alpha]_C(d), [\beta]_C(d)\}$ as a tight upper bound for conjunction, but also as a tight lower bound for disjunction.

Now assume that both sets are as disjoint as possible. We distinguish:

Case 1. $[\alpha]_C(d) + [\beta]_C(d) \leq 1$: Then $\alpha(d)(C) \cap \beta(d)(C) = \{\}$, thus $[A \,and\, B]_C(d) = 0$ and $[A \,or\, B]_C(d) = [\alpha]_C(d) + [\beta]_C(d)$.

Case 2. $[\alpha]_C(d) + [\beta]_C(d) > 1$: Then $\alpha(d)(C) \cap \beta(d)(C) \neq \{\}$. As we assume the sets to be as disjoint as possible, their intersection is as small as possible; therefore $|\alpha(d)(C) \cap \beta(d)(C)| = [\alpha]_C(d) + [\beta]_C(d) - 1$, and $\alpha(d)(C) \cup \beta(d)(C) = 1$

Combining the cases yields the specified bounds. □

Remark. Note that $*_{\mathrm{G}} = \min$ and $\bar{*}_{\mathrm{G}} = \max$ are the Gödel t-norm and co-t-norm, respectively. Moreover, $*_{\mathrm{L}} = \lambda x, y.\, \max\{0, x + y - 1\}$ and $\bar{*}_{\mathrm{L}} = \lambda x, y.\, \min\{1, x + y\}$ are the Łukasiewicz t-norm and co-t-norm, respectively. In other words, Theorem 1 shows that the truth functions of (strong) conjunction and (strong) disjunction in Gödel and Łukasiewicz logic (see [11]) correspond to opposite extremal cases of context based evaluations of conjunction and disjunction.

The above analysis on logical predicate operators can easily be lifted to the propositional level. For a sentence like *Jana is tall* its meaning $[\![Jana\ is\ tall]\!]$ is a filter (rather than an element filter). Usual logical connectives on propositions can be defined in analogy to Definition 2:

Definition 3.

- $[\![\phi \wedge \psi]\!] =_{df} \lambda C.[\![\phi]\!](C) \cap [\![\psi]\!](C)$
- $[\![\phi \vee \psi]\!] =_{df} \lambda C.[\![\phi]\!](C) \cup [\![\psi]\!](C)$
- $[\![\neg\phi]\!] =_{df} \lambda C.C \backslash [\![\phi]\!](C)$

In the following the set of all propositions formed in this way is called Prop.

Similarly to the predicate level, we can associate a 'degree of truth' $\|\phi\|_C$ for every $\phi \in$ Prop by applying the filter $[\![\phi]\!]$ to context C:

$$\|\phi\|_C =_{df} \frac{|[\![\phi]\!](C)|}{|C|}.$$

In other words we identify the degree of truth of ϕ in a context C with the fraction of worlds in C that survive the update with the filter $[\![\phi]\!]$. E.g., returning to the context C specified in the example following Definition 2, *Jana is tall* is true to degree 3/5 in C since three out of five worlds in C classify *Jana*'s height as above the relevant local standard of tallness.

Once more we note that contexts allow to model specific constraints on the worlds (i.e. contextually relevant possible precisifications) of which they consist. Therefore, in general, there are no truth functions that determine $\|\phi \wedge \psi\|_C$ and $\|\phi \vee \psi\|_C$ in terms of $\|\phi\|_C$ and $\|\psi\|_C$ alone. However the optimal bounds of Theorem 1 also apply at the level of sentences:

- $*_{\mathrm{L}}(\|\phi\|_C, \|\psi\|_C) \leq \|\phi \wedge \psi\|_C \leq *_{\mathrm{G}}(\|\phi\|_C, \|\psi\|_C)$, and
- $\bar{*}_{\mathrm{G}}(\|\phi\|_C, \|\psi\|_C) \leq \|\phi \vee \psi\|_C \leq \bar{*}_{\mathrm{L}}(\|\phi\|_C, \|\psi\|_C)$,

where $*_{\mathrm{G}}(\bar{*}_{\mathrm{G}})$ and $*_{\mathrm{L}}(\bar{*}_{\mathrm{L}})$ are the Gödel and Łukasiewicz t-norms (co-t-norms), respectively.

4 Translating voting semantics to contexts

As we have seen in Section 3, the context based semantics of logical connectives is more fine grained than any specification by some particular truth function over degrees. The fraction of worlds surviving an update with $[\![\phi \wedge \psi]\!]$ is not determined by the fractions of worlds surviving the filters $[\![\phi]\!]$ and $[\![\psi]\!]$, respectively: t-norm based truth functions provide optimal bounds, but in general the internal structure of contexts determines the corresponding fractions of worlds surviving updates with logically complex propositions. The following question arises: can one constrain and/or modify the structure of contexts in a manner that leads to standard fuzzy truth functions at the level of such contexts. For a positive answer we rely on an analogy between Lawry's *voting semantics* [19] and our (or rather Barker's) version of contextual semantics.

To explain the assignment of truth values $\in [0, 1]$ to a statement ϕ Lawry [19], but also many other researchers (e.g., [7, 13]) suggest to consider the following scenario. Ask each of N agents whether she accepts the statement ϕ. It is assumed that the agents are all competent speakers of the respective language and are fully informed about the relevant facts. Therefore they will all agree on whether ϕ is to be accepted or to be rejected if ϕ is a precise statement. However, if ϕ is vague[9] then they may diverge on their judgements in spite of their linguistic competence and factual knowledge. In this setting one assigns the 'truth value' $v = n/N$ to ϕ, where n is the number of agents that accept ϕ.

Let us write $a_s(\phi) = 1$ if agent s accepts ϕ and $a_s(\phi) = 0$ otherwise. If the agents have to satisfy the following consistency conditions

$$\begin{aligned} a_s(\phi \wedge \psi) = 1 &\iff a_s(\phi) = 1 \text{ and } a_s(\psi) = 1 \\ a_s(\phi \vee \psi) = 1 &\iff a_s(\phi) = 1 \text{ or } a_s(\psi) = 1 \\ a_s(\neg\phi) = 1 &\iff a_s(\phi) = 0 \end{aligned}$$

then the resulting global 'fuzzy truth value assignment' turns out to be simply a probability function (see, e.g., [21]) and therefore does not justify a truth functional semantics of fuzzy logic if the agents' votes are independent. However, if

[9] We deliberately focus on vagueness and ignore other forms of indeterminateness and uncertainty here.

we require that the agent's voting behaviour is determined by an associated 'degree of scepticism' in a particular way, than usual fuzzy truth functions emerge.

Definition 4. *A family of functions* $a_\sigma : \mathrm{Prop} \mapsto \{0,1\}$*, where* $\sigma \in [0,1]$ *is called a* scepticism degree based voting behaviour *if the following conditions hold:*

$$\begin{array}{rcl} \text{if } \sigma \leq \sigma' \text{ and } a_\sigma(\phi) = 0 \text{ then } a_{\sigma'}(\phi) = 0 \\ a_\sigma(\phi \wedge \psi) = 1 & \Longleftrightarrow & a_\sigma(\phi) = 1 \text{ and } a_\sigma(\psi) = 1 \\ a_\sigma(\phi \vee \psi) = 1 & \Longleftrightarrow & a_\sigma(\phi) = 1 \text{ or } a_\sigma(\psi) = 1 \\ a_\sigma(\neg\phi) = 1 & \Longleftrightarrow & a_{1-\sigma}(\phi) = 0 \end{array}$$

The intended interpretation of the *scepticism degree* σ is the level of willingness to assert a positive statement. The first condition means that an agent rejects at least all those propositions that are rejected by less skeptic agents. The condition for negated statements implies that an agent with a high degree of scepticism is willing to accept $\neg\phi$ whenever an agent with inverted (low) degree of scepticism is willing to reject ϕ. This implies that, in general, agents do not evaluate classically: we may have $a_\sigma(\phi \vee \neg\phi) = 0$ but also $a_\sigma(\phi \wedge \neg\phi) = 1$; only $a_{0.5}$ is always a classic valuation. To obtain a (global) fuzzy valuation from such families of (local) para-consistent $\{0,1\}$-valuations, we have to measure 'amounts of acceptance'.

Definition 5. *Let* $\Lambda = \{a_\sigma : \sigma \in [0,1]\}$ *be a scepticism degree based voting behaviour and let* μ *be a measure on the Borel subsets of* $[0,1]$*. Then the corresponding fuzzy truth value assignment is defined by*

$$v^\mu_A(\phi) = \mu\{\sigma \in [0,1] : a_\sigma(\phi) = 1\}$$

Proposition 1. ([19]) *For all scepticism degree based voting behaviors* Λ *and measures* μ*, as above, we have:*

$$\begin{array}{rcl} v^\mu_A(\phi \wedge \psi) & = & \min(v^\mu_A(\phi), v^\mu_A(\psi)) \\ v^\mu_A(\phi \vee \psi) & = & \max(v^\mu_A(\phi), v^\mu_A(\psi)) \end{array}$$

Moreover, if μ *is symmetric, i.e. if* $\mu[a,b] = \mu[1-b, 1-a]$ *for* $0 \leq a \leq b \leq 1$*, then*

$$v^\mu_A(\neg\phi) \quad = \quad 1 - v^\mu_A(\phi)$$

How does this relate to contextual dynamic semantics? The most obvious transfer of voting semantics to contexts is to associate with each world a value that directly corresponds to the scepticism degree of an agent and to evaluate logically complex statements as specified above. But remember that this entails that local evaluations violate either the law of excluded middle ($\phi \vee \neg\phi$) or the law of contradiction ($\neg(\phi \wedge \neg\phi)$) in general. Of course, a world c of a context C is something different than a voter among many voting agents. But c can be viewed as a *local semantic test*: it specifies for each sentence ϕ whether ϕ holds according to certain precisified standards or not. It does not seem to be unnatural to compare these semantic tests with respect to their *strictness* in analogy to the comparison of agents with respect to degrees of scepticism. Moreover, considering the intended application of contextual semantics, we may assume that only one or at most a few directly related predicates are relevant in a given context. Also the domain of any particular context can realistically

be assumed to be small. This makes it plausible that worlds of a context may often be characterized solely by their *degree of strictness*. Let us illustrate this by an example from natural language. A realistic context for evaluating

(1) *Jana is tall*

might be represented by worlds (i.e. precisifications) that agree on Jana's actual height (say 178cm) but differ in their standards of accepting 178cm as being above the local standard of tallness. Obviously we can then define linearly ordered degrees of strictness induced by increasing standards of tallness for the worlds of such a context. A similar observation holds for

(2) *The weather is cold today*

Again, we have no troubles to extract degrees of strictness corresponding to decreasing threshold values (temperatures) for accepting (2). Using our re-interpretation of voting semantics we can extract truth values $\in [0, 1]$ for (1) and for (2) in the respective contexts indicated above. In contrast, one might argue that there simply is *no natural context* in which the *conjunction* of (1) and (2) has to evaluated, which nicely fits our model.

While the above remarks may be sufficient to justify the focus on contexts with associated linearly ordered degrees of strictness, the fact that the translation of voting semantics to contexts calls for 'non-classical worlds' seems to be more problematic. However we claim that this is compatible with Barker's context based model [1], as introduced in Section 2. Note that Barker does not provide a semantics for logical connectives. Only non-compound vague predicates and vagueness-related predicate modifiers are investigated. While, following voting semantics, one can straightforwardly generalize to include *conjunctions* and *disjunctions* at the local level of individual worlds, *negation* is viewed in this model as an inherently global operator, which only receives meaning at the level of whole contexts.

5 Summary and outlook

We started by noting the fact that linguists usually analyze the semantics of vague words by reference to contexts of utterance that register relevant possible precisifications. This seems to be at variance with the degree based approach to vagueness suggested by fuzzy logic. However, taking Barker's [1] version of dynamic (update) semantics as a concrete point of reference, we have demonstrated that fuzzy sets can be associated in a systematic manner with contexts and corresponding filters as used in Barker's model. While the structure of context filters used to specify the different meanings of modifiers like *very*, *definitely*, and *clearly* allows to take into account information that is abstracted away in corresponding fuzzy sets, standard t-norm based operators faithfully register the extremal cases that may result from applying logical connectives to vague predicates and sentences.

While it is rather straightforward to identify intermediate truth values with the fraction of worlds in a given context that survive certain updates codifying the meaning of vague expressions, it is not clear how one might derive specific truth functions in such a setting (beyond providing the indicated bounds). This problem, of course, is just a particular instance of a well known challenge for

deductive fuzzy logic: how to justify particular truth functions with respect to more fundamental semantic notions, like votes or arguments for and against accepting a vague assertion. In [23] Jeff Paris provides an overview over semantic frameworks for fuzzy logics that support truth functionality. Here we picked a particular approach, namely so-called voting semantics as suggested by Lawry [19] to illustrate how one might connect context based update semantics with frameworks that model the meaning of logical connectives by particular t-norm based truth functions.

We emphasize that both, Barker's specific update functions over contexts and Lawry's voting semantics, should be understood as just two particular spots on either side of the river dividing linguistics from fuzzy logic, that may be chosen as end points of a bridge crossing that troubled water. On the linguistic side context and precisification based approaches suggested, e.g., by Kennedy [16], Kyburg and Moreau [18], and already earlier by Pinkal [25] and Bosch [3] are certainly worth investigating from this perspective. On the fuzzy logic side we just mention similarity semantics [27, 17, 30], Robin Giles's dialogue and betting game based characterization of Łukasiewicz logic [9, 8] (extended to other logics in [4, 6]), acceptability semantics [22], rerandomising semantics [13, 11], and approximation semantics [2, 23] as alternative candidates for corresponding bridge heads. We plan to explore at least some of these options in future work. In any case, we hope to have shown already here that constructing such a bridge is neither a futile nor a completely trivial matter.

References

[1] C. Barker. The dynamics of vagueness. *Linguistics and Philosophy*, 25(1):1–36, 2002.

[2] A.D.C. Bennett, J.B. Paris, and A. Vencovska. A new criterion for comparing fuzzy logics for uncertain reasoning. *Journal of Logic, Language and Information*, 9(1):31–63, 2000.

[3] P. Bosch. "Vagueness" is context-dependence. A solution to the sorites paradox. *Approaching Vagueness*, pages 189–210, 1983.

[4] A. Ciabattoni, C.G. Fermüller, and G. Metcalfe. Uniform rules and dialogue games for fuzzy logics. *Lecture Notes in Computer Science*, 3452:496–510, 2005.

[5] C.G. Fermüller. Theories of vagueness versus fuzzy logic: can logicians learn from philosophers? *Neural Network World*, 13(5):455–466, 2003.

[6] C.G. Fermüller. Revisiting Giles's Game—Reconciling Fuzzy Logic and Supervaluation. In *Selected Papers from the Prague Colloquium on Logic, Games, and Philosophy: Foundational Perspectives*. Springer, 2004.

[7] B.R. Gaines. Foundations of fuzzy reasoning. *International Journal of Man-Machine Studies*, 8(6):623–668, 1976.

[8] R. Giles. Foundations for quantum mechanics. *Journal of Mathematical Physics*, 11(7):2139–2161, July 1970.

[9] R. Giles. A non-classical logic for physics. *Studia Logica*, 33(4):397–415, 1974.

[10] J. Groenendijk and M. Stokhof. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100, 1991.

[11] P. Hajek. *Metamathematics of fuzzy logic*. Kluwer academic publishers, 2001.

[12] I. Heim and A. Kratzer. *Semantics in generative grammar*. Blackwell Publishers, 1998.

[13] E. Hisdal. Are grades of membership probabilities? *Fuzzy Sets and Systems*, 25(3):325–348, 1988.

[14] R. Keefe. *Theories of vagueness*. Cambridge University Press, 2000.

[15] R. Keefe and P. Smith, editors. *Vagueness: A reader*. MIT press, 1999.

[16] C. Kennedy. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45, 2007.

[17] R. Kruse, J.E. Gebhardt, and F. Klowon. *Foundations of fuzzy systems*. John Wiley & Sons, Inc. New York, NY, USA, 1994.

[18] A. Kyburg and M. Morreau. Fitting words: Vague language in context. *Linguistics and Philosophy*, 23(6):577–597, 2000.

[19] J. Lawry. A voting mechanism for fuzzy logic. *International Journal of Approximate Reasoning*, 19(3-4):315–333, 1998.

[20] D. Lewis. General semantics. *Synthese*, 22(1):18–67, 1970.

[21] J.B. Paris. *The uncertain reasoner's companion: a mathematical perspective*. Cambridge Univ Press, 1994.

[22] J.B. Paris. A semantics for fuzzy logic. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 1(3):143–147, 1997.

[23] J.B. Paris. Semantics for fuzzy logic supporting truth functionality. pages 82–104, 2000.

[24] B.H. Partee. Montague semantics. *Handbook of Logic and Language*, pages 5–91, 1997.

[25] M. Pinkal. *Logic and lexicon*. Kluwer Academic Publishers Boston, 1995.

[26] R. Rooij. Vagueness and linguistics. *The Vagueness Handbook, to appear*.

[27] E. Ruspini. The semantics of fuzzy logic. In *NASA, Lyndon B. Johnson Space Center, Proceedings of the Second Joint Technology Worshop on Neural Networks and Fuzzy Logic,*, volume 2, 1991.

[28] S. Shapiro. *Vagueness in context*. Oxford University Press, 2006.

[29] R. Stalnaker. On the representation of context. *Journal of Logic, Language and Information*, 7(1):3–19, 1998.

[30] E. Trillas and L. Valverde. An inquiry into indistinguishability operators. *Aspects of vagueness*, pages 231–256, 1984.

[31] T. Williamson. *Vagueness*. Burns & Oates, 1994.

How People Interpret an Uncertain IF*

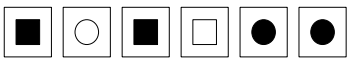
Andrew J. B. Fugard†, Niki Pfeifer,
Bastian Mayerhofer, and Gernot D. Kleiter
Department of Psychology, University of Salzburg, Austria

Abstract

Conditionals are central to inference. Before people can draw inferences about a natural language conditional, they must interpret its meaning. We investigated interpretation of uncertain conditionals using a probabilistic truth table task, focussing on (i) conditional event, (ii) material conditional, and (iii) conjunction interpretations. The order of object (shape) and feature (color) in each conditional's antecedent and consequent was varied between participants. The conditional event was the dominant interpretation, followed by conjunction, and took longer to process than conjunction (mean difference 500 ms). Material conditional responses were rare. The proportion of conditional event responses increased from around 40% at the beginning of the task to nearly 80% at the end, with 55% of participants showing a qualitative shift of interpretation. Shifts to the conditional event occurred later in the feature-object order than in the object-feature order. We discuss the results in terms of insight and suggest implications for theories of interpretation.

1 Introduction

Consider a fair die with the following patterns on the sides:



The die is thrown and lands with one side facing up. How sure can you be that *if the side shows a square, then the side shows black*? Before you can respond you first interpret the meaning of the conditional and the task you are meant to perform. If your answer was 2/3, then it is likely you interpreted the conditional as the conditional event; if your answer was 5/6, then it is likely your interpretation was the material conditional of classical logic; and if your answer was 2/6 (or 1/3), then it is likely your interpretation was the conjunction. The present contribution investigates how people interpret the indicative conditional using this dice task. Specifically we aim to investigate: (i) what are the dominant

*Supported by the European Science Foundation EUROCORES programme LogICCC, and the Austrian Science Fund projects I141 and P20209. Thanks to Hans Lechner for producing our response box and Sabine Eichbauer for response sheet design and scanning.

† *Corresponding author.* Email: andy.fugard@sbg.ac.at

interpretations of the conditional in the task, (ii) how do linguistic features influence interpretation, and (iii) does interpretation change over time?

The conditional in all its forms has received much attention across the disciplines as it is ubiquitous in inference, for instance in conversation and problem solving, and in inferences about inference, for instance in mathematical logic. Until the late 1990s, the majority of psychological theories of conditional reasoning have used classical logic as the framework for competence and performance models. For instance the theory of *mental models* stems from a fragment of model theory of classical logic [5]. The *mental rules* or *mental logic* theories stem from Gentzen's natural deduction systems for classical logic (e.g., [14]). The conjunction ($A \wedge B$) and the material conditional ($A \Rightarrow B$) interpretations of the natural language conditional are postulated by the mental model theory [5], which is one of the most influential theories in the psychology of reasoning.

An alternative view gaining in popularity in psychology is that the indicative 'if A, then B' is interpreted as a conditional event, $B|A$ [4, 8, 11]. Ramsey [13, p. 155] argued that when people infer their degree of belief in 'if A, then B', they assume A, and 'fix their degrees of belief' in B. If the antecedent A turns out to be false, 'these degrees of belief are rendered *void*'. Unlike the conditional event, the material conditional can be reduced in various ways to combinations of Boolean operators: for instance $A \Rightarrow B$ is equivalent to $\neg A \vee B$, which is a disjunction, and to $\neg(A \wedge \neg B)$, a negated conjunction. A psychological implementation of the 'Ramsey test' has been proposed [4, p. 325]. In this, conjunction responses are argued to be due to an incomplete execution of the test, because of limited working memory or insufficient motivation.

Given the variety of interpretations shown on reasoning tasks, it has been proposed to separate reasoning *to* interpretations and *from* interpretations [15]. Reasoning to an interpretation requires (i) a formal language to be chosen, (ii) a semantics to be assigned, and (iii) a characterization of when an argument is valid [15, p. 25]. Once these choices have been made, then reasoning from the fixed interpretation (i.e., derivation) may proceed. From this viewpoint, errors in reasoning may be due either to mismatches in interpretation (e.g., between experimenter and participant) or a failure of derivational processes. For the above dice task we assume that the language, semantics, and derivational apparatus of a probability theory are appropriate. While there are many approaches to probability, we favor *coherence based probability logic* [3]. Coherence has many advantages for psychological modeling compared to alternative approaches [10, 11], e.g., conditional events are *primitive* and not defined by unconditional probabilities, they are *undetermined* if the antecedent is false, and probabilities are conceived as *degrees of belief* rather than 'objective' quantities [3]. The main interpretational problem in the dice task is deciding whether the natural language 'if A, then B' is interpreted as (i) a conditional event ($B|A$), (ii) a conjunction ($A \wedge B$), or as (iii) a material conditional ($A \Rightarrow B$).

The standard test paradigm for investigating how people interpret indicative uncertain conditionals is the *probabilistic truth table task* [4, 9]. In this task, the joint frequency distribution is provided (i.e., frequencies of conjunctions) and participants are asked to assess how sure they are that a conditional is true. A characteristic feature of probabilistic truth table tasks is that they are problems under full probabilistic knowledge, since all joint probabilities are given. Full probabilistic knowledge allows for precise (i.e., point rather than interval) probability assessments of the conclusions and the coherent predictions according to

the different interpretations are easily calculated. The task allows the experimenter to infer how the participants interpret the conditional. Overall studies using probabilistic truth table tasks have found that just over half of participants responded with the conditional event interpretation and the remainder responded with a conjunction interpretation [4, 9]. Little support was observed for the material conditional interpretation.

The present study extends previous research on uncertain conditionals by (i) presenting the task material graphically, without using numerals; (ii) not priming a representations in terms of joint frequencies; (iii) presenting a series of systematically enumerated items; (iv) studying reaction times; (v) investigating the time-course of interpretation within-participants, for instance whether there are any shifts of interpretation; and (vi) studying facilitation effects of object-first versus feature-first conditionals.

Task development We developed a task concerning six-sided dice, using patterns on each side of a given die rather than the usual numerals. These patterns were varied systematically from two independent dimensions: shape (e.g., square or circle) and the shape's color (e.g., red or blue). There are 84 possible assignments of two shapes and two colors to the six-sides of the dice. *A priori*, many of the resulting items do not distinguish between interpretations, however this is not always a problem if few or no participants give a non-uniquely classifiable response for a particular item, e.g, if an item does not distinguish between $P(B|A)$ and $P(A \Rightarrow B)$, but no participants give this non-distinguishable response but rather respond with $P(A \wedge B)$. The non-uniquely classifiable responses may still help to exclude certain interpretations, e.g., if a response can only be classified as either $P(B|A)$ or $P(A \wedge B)$, it excludes $P(A \Rightarrow B)$. From each participant's pattern of responses we can infer how they interpreted the conditional.

The task and instructions were implemented in Python using the Pygame graphical library.[1] Participants were told that the aim of the experiment was to investigate how people understand if-then sentences. It was emphasized that the die varied between trials and that they were to reason about each independently. Three examples were also given of how the sides of the die would be represented on screen. A simple animation was shown to convey the idea of a die being placed in a cup, randomly shaken, and then the cup placed on the table so that one cannot see what side of the die shows up. Four example trials were then presented to check that the participant understood the response modality. These asked how sure the participant can be that atomic sentences hold, e.g., 'The side shows a circle' (*Die Seite zeigt einen Kreis*).

Each test trial began with a fixation cross displayed for 1 second. Participants were shown the patterns on the sides of the die and were asked to estimate how sure they were that a given conditional, for example, 'If the side shows a square, then the side shows red' (*Wenn die Seite ein Viereck zeigt, dann zeigt die Seite rot*), was true[2] of a thrown die. Since we were interested in studying interpretation rather than mental arithmetic, we asked participants to respond with 'x out of y' (x *aus* y), rather than a probability or percentage, thus elimi-

[1] Python version 2.6.1 (`www.python.org`) and Pygame version 1.8.1 (`www.pygame.org`).

[2] The German word *stimmt* was used which is weaker than the German word for 'true' (*wahr*).

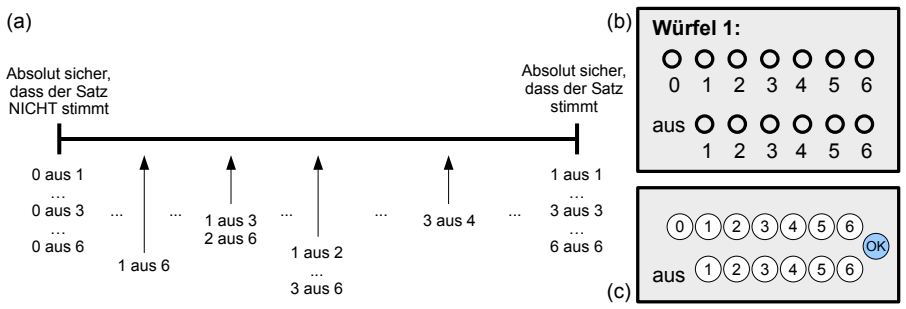


Figure 1: (a) Diagram used to convey the meaning of 'out of'. (b) Example of item response format on answer sheet. (c) Diagram of response box. '*Würfel*' translates to 'die'; '*aus*' to 'out of'; and '*Absolut sicher, dass der Satz (NICHT) stimmt*' to 'Absolutely certain that the sentence is (NOT) true'.

nating the need to divide numbers and rescale, which many people find difficult. We presented a visual scale to explain the meaning of 'out of' (see Figure 1(a)) and showed that the numerator should not exceed the denominator.

The task was piloted in a seminar room to 18 students, with presentation using a data projector. We selected 77 items such that the probability of the antecedent is not zero, so that the conditional event is determined. From the original 77 items, responses to 51 could all be uniquely classified. Counting first each participant's most common strategy, 13 responded mostly with the conditional event (median 49, range 10–51), four with the conjunction (median 41, range 36–50), and one person according to none of the competence models. There were no responses according to the material conditional. Feedback from students was used to improve instructions for Experiment 1.

First steps towards a process model All reasoning tasks involve premises and a conclusion, but what exactly are they in this task? The instructions are supposed to communicate that the die is six sided, fair and thrown randomly, and that the probability of a side landing up is 1/6. Probabilities are obtained by counting the relevant joint or marginal frequencies (i.e., the frequencies of the conjuncts). The conclusion is a natural language conditional and must be interpreted. Table 1 shows how the chosen interpretation determines which premises are relevant, and how the presented information may be used to compute the coherent probability inferences for the three predicted interpretations. These choices of premises and how they are integrated are not unique. For example the probability of the material conditional interpretation may be calculated using only one joint frequency, $P(A \Rightarrow B) = \frac{6-|A \wedge \neg B|}{6}$ (where $|\varphi|$ denotes the frequency of φ), rather than summing up three joint frequencies. Also $|A|$ may be inferred from the sum $|A \wedge B| + |A \wedge \neg B|$.

Although mathematically the task is straightforward, the psychological processes required to solve the task are complex. To understand possible processes we must first decompose the task into the abilities required for its solution. The conditional, 'if A, then B', must be parsed and committed to working memory. Previous experiments on generating analogies suggest that the order of the object-feature positions in the conditionals may affect performance [6]. For a

$\lvert A \wedge B \rvert = f_1$	$\lvert A \wedge B \rvert = f_1$	$\lvert A \wedge B \rvert = f_1$
$\lvert A \rvert = f_2$	$\lvert Sides \rvert = 6$	$\lvert \neg A \wedge B \rvert = f_2$
$\models P(B \lvert A) = f_1/f_2$	$\models P(A \wedge B) = f_1/6$	$\lvert \neg A \wedge \neg B \rvert = f_3$
		$\lvert Sides \rvert = 6$
		$\models P(A \Rightarrow B) = \frac{f_1+f_2+f_3}{6}$

Table 1: Examples of premises obtainable from the dice presentations and how they may be used to infer the probability of the if-then according to the three interpretations.

feature, F (e.g., red), and concrete object, o (e.g., a square), it is easier first to form a representation of o, and second bind it to $F(o)$, than first to form a representation of F, and second bind it to $F(o)$. For the conditional event and material conditional interpretations (though not conjunction), order matters, thus this must be respected in the memory representation of the conditional. The visual depiction of the sides of the die must be perceived and categorized. Runs of patterns of the same type may facilitate this process. For each of the competence models, the number of sides with each relevant property (relative to interpretation) must be counted. For instance for the conditional event interpretation, participants need $|A \wedge B|$ and $|A|$. There are different ways of obtaining these frequencies. One may start at the left-most die-side and count $y = |A|$ and then count how many of these also had the property B; denote the result x. Then the response is 'x out of y'. Alternatively one may begin by counting $x = |A \wedge B|$, store the value, and then count $y = |A|$, responding 'x out of y'. In both cases the result will be the same. At each point in the task it is possible to refresh one component, e.g., the number of sides with a particular property may be recounted or the conditional statement re-parsed. An additional memory component is required for goal maintenance, e.g., remembering not only the conditional and counts, but also the very fact that these have been remembered, what information has to be obtained next from the task presentation, and how the information must be integrated. Finally the response has to be made.

This sketch allows us to generate experimental hypotheses which may be operationalized. In Experiment 1 we tested the interpretation of the conditional. Participants may respond with a conjunction probability if they leave out the step of computing $|A|$, e.g., because of a failure of goal memory. In Experiment 2 we test if the order of responses reveal the strategy pursued. Participant who first calculate $|A|$ may wish to unburden their working memory before calculating $|A \wedge B|$. Allowing them to do so may reveal their order of processing. Reaction times ought to be faster for a conjunction rather than a conditional event interpretation (as you need not count both $|A \wedge B|$ and $|A|$). In both experiments we tested for an effect of the object-feature (i.e., shape-color) order in the conditional using a between-participant design. Previous work found a reaction time benefit for the object-feature order, but we also investigated whether the proportion of conditional event responses was influenced by order.

2 Experiment 1

Method The task was presented in a lecture theater to 66 students (57 females and 9 males), whose ages ranged from 20 to 40 ($M = 23.8$; $SD = 3.5$), at the be-

ginning of an introductory psychology course to thinking and reasoning (before conditional reasoning had been introduced) at the University of Salzburg.

For the between-participant manipulation of object-feature order, 33 participants were assigned to the object-feature condition, and 33 were assigned to the feature-object condition (conditions alternated in the distribution of booklets). From the original bank of 84 items, 71 were selected such that probability of the antecedents for both object-feature orders were not zero. The instructions and item presentation were computer controlled and displayed on the theater screen using a data projector. Responses were given on a response sheet designed for automatic scoring (see Figure 1(b)). The item number was displayed on screen and on the response sheet. For the first trial, participants were given 30 seconds to respond. The second trial lasted 10 seconds, followed by a pause during which the experimenter explained that the task was about to begin. Each test trial lasted 10 seconds, the end of which was indicated by three beeps.

Results and discussion Responses to 46 of the 71 items could be uniquely classified. No effect was found for the object-feature order, so we pooled the data of both conditions. Counting each participant's modal response type, 50 participants responded mostly with the conditional event (median 43, range 15–46), eight with the conjunction (median 27, range 17–46), and six with some other non-predicted response (median 27, range 23–34). There was one participant responding mostly with the reversed conditional event (a score of 23) and one material conditional responder (all responses).

As participants proceeded through the task, the proportion of conditional event interpretations increased ($r(44) = .82$, $p < .001$) and the proportion of conjunction responses decreased ($r(44) = -.73$, $p < .001$). See Figure 2(a–c). We have two explanations for the convergence on the conditional event. One is that participants learn the conditional event interpretation as they progress through the task. It could be that after many presentations of the dice stimuli, the antecedent frequencies become more salient and are included in the interpretation of the probability of the conditional. Another explanation is in terms of speed-accuracy trade-off. More time may be required to process the material using the conditional event interpretation, so those participants who appeared to shift interpretation actually had a fixed interpretation, but adapted to task demands. Those who shifted from a conjunction response may first have calculated the joint probability. The absence of an effect for the object-feature order may be because the conditional remained constant throughout the task and thus needed to be processed only once. This problem will be addressed in the next experiment.

3 Experiment 2

In this experiment we adapted the task for computer-controlled individual testing to (i) collect response times, (ii) determine whether participants respond first with the numerator or with the denominator, (iii) vary the shapes and colors in the conditionals between trials to ensure reprocessing of the conditional for each item, and (iv) improve experimental conditions compared to those in a lecture theater. We hypothesized that response times will be shorter for participants using a conjunction interpretation as they have to count only one joint and no

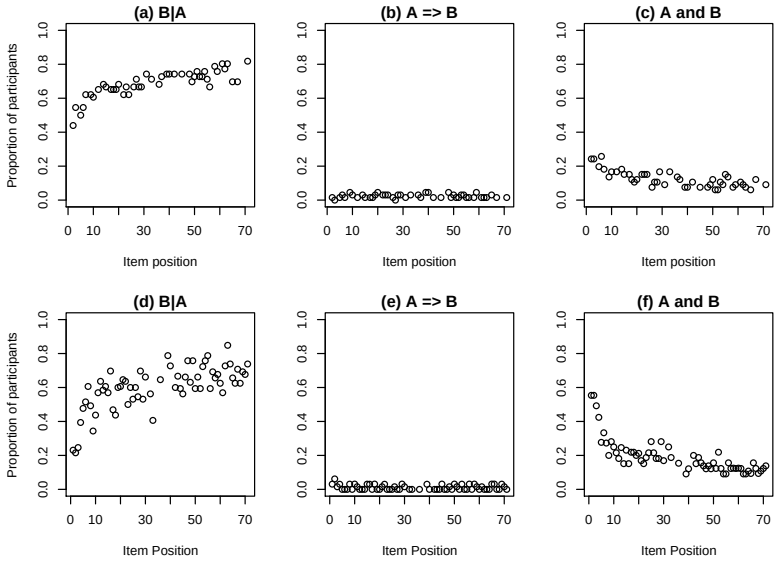


Figure 2: Proportion of participants giving a response of each class, as a function of item position in (a–c) Experiment 1 and in (d–f) Experiment 2. Only the uniquely classifiable items are included.

marginal frequency. Further we hypothesize that if the object is presented in the antecedent, then participants will be faster in evaluating its probability, than if it is presented in the consequent.

Method Participants were 65 students (32 females and 33 males) whose ages ranged from 18 to 30 ($M = 22.9$; $SD = 2.9$) from the University of Salzburg, 49 of whom study a natural science, and 16 study a humanities subject. Students of psychology, mathematics, or with a special background in formal logic, were not included in the sample. We paid 5 Euros for participation.

A button box was designed (see Figure 1(c)) with a layout similar to the pen-and-paper response sheet layout used in Experiment 1. We added an extra shape (triangle) and color (green), and randomly cycled through colors and shapes to encourage participants to reprocess the conditional, thus making it more likely that an effect of object-feature order can be detected. The areas of the objects were adjusted so that they have the same perceivable area. Between-participant we crossed sex, random order (one order, forwards/backwards), and object-feature order. Within-participant we varied the frequencies of shapes and colors with the constraint that the probabilities of the antecedents are not zero. Each item remained on screen until participants made their responses.

Results and discussion Counting each participant's modal response type for the 46 uniquely classifiable items, 45 participants responded mostly with the conditional event (median 40, range 19–46), 11 with the conjunction (median 42, range 20–46), 2 with the reversed conditional event (18 and 39), nobody with the material conditional, and 7 with some other response (median 29, range 19–37). We replicated the result found in Experiment 1: as participants proceeded through the task, the proportion of conditional event interpretations increased

($r(63) = .68$, $p < .001$) and the proportion of conjunction responses decreased ($r(63) = -.73$, $p < .001$).[3] See Figure 2(d–f).

We sought to investigate within-participants the nature of this increase in conditional event responses. Do participants smoothly increase the probability of a conditional event interpretation, or is there a sudden shift in interpretation? Visual inspection of responses suggested that many participants shifted suddenly to a particular interpretation after some time. Thus we decided to investigate interpretation shifts systematically to detect for whom and when this occurred. To find a shift point for each participant, we used the following simple algorithm:

1. Let $S = \langle s_1, \ldots s_{71} \rangle$ denote the binary sequence of 71 conditional event scores. $C = \langle c_1, \ldots c_{71} \rangle$ denotes a sequence of 71 scores, where each element of C represents how many different interpretations a '1' in the conditional event score could represent, e.g., if the ith response could be either a conditional event or conjunction, then $s_i = 1$ and $c_i = 2$. For a given i, $c_i \in [0, 5]$ (0 if the response is an 'other' response).

2. Use these two sequences to create a weighted sequence, $W = \langle w_1, \ldots w_{71} \rangle$: if $c_i = 0$, then set $w_i := 0$, as this response is an 'other' response; otherwise set $w_i := s_i / c_i$.

3. For every $i \in [2, 71]$, compute the proportions $l_i = \sum_{j=1}^{i-1} w_j/(i-1)$ and $r_i = \sum_{j=i}^{71} w_j/(71 - i + 1)$. Note that r_i includes position i.

4. The split point is found by maximizing $r_i - l_i$. When there is more than one i where this difference is maximal, we take the first.

We also computed the modal interpretation to the left and right of this split point, and the proportion of responses of these modal types, using the 46 uniquely classifiable responses. Just over half of the participants (36, around 55%) shifted from some other interpretation to the conditional event interpretation. Of these, the majority (29, around 80%) shifted from the conjunction interpretation, three from the reversed conditional event ($A|B$), three from some non-classifiable response, and only one from the material conditional—but in it's reversed form ($B \Rightarrow A$).

The earliest shift occurred at item position 2 (one participant), with most (64%) shifting at least by position 8. Figure 3(a) shows the distribution of the splits. Figures 3(b) and (c) show the proportion of responses of the modal type to the left and conditional event to the right of (and including) the split. As may be seen, most participants are very consistent once they have shifted to the conditional event (mean proportion of conditional event responses after the shift is .93, $SD = .10$).

We also have some self-report data from the participants on their strategies. Participant 34 (who settled into a conjunction interpretation) said: 'I only looked at the shape and the color, and then always out of 6; this was the quickest way.' Participant 37, who shifted from the conjunction to the conditional event, said: 'In the beginning [I] always [responded] 'out of 6', but then somewhere in

[3]Correlations were computed using the original item positions (1 to 71), not their relative position (1 to 46). Since data from the two random orders were pooled, 65 rather than 46 pairs of values resulted, as data were available for a particular item position in only one direction for 19 positions.

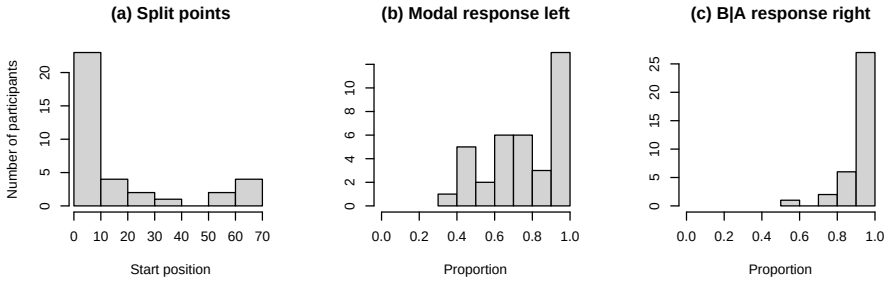


Figure 3: (a) Distribution of split point positions. (b) Proportion of responses to the left of the split point which are of the modal class to the left. (c) Proportion of responses to the right of (and including) the split point which are consistent with the conditional event interpretation. (For the 36 participants who shift.)

the middle... Ah! It clicked and I got it. I was angry with myself that I was so stupid before.' Five participants spontaneously reported when they shifted during the task, e.g., saying, 'Ah, this is how it works.' Such unprompted comments are typical indicators of insight effects [2].

Fourteen participants (around 20%) pressed a button from the bottom row first at least once. Eight of these did so exactly once, and the remainder between 10–40 times out of 71 responses. Only one conjunction response from one participant was made by pushing the bottom button first. For conditional event responses, only four participants pressed the bottom button first a non-negligible number of times: 15–25. Thus the hypothesized benefit of unburdening working memory has not received strong support.

We tested our hypothesis that participants would be faster for a conjunction versus a conditional event response using mixed-effects models[4] with the basic structure as follows:

$$\begin{aligned} log(RT_{ip}) &= \beta_0 + \gamma_{0p} + \gamma_{1i} + \beta_1 \cdot pos_{ip} + \beta_2 \cdot pos_{ip}^2 \\ &+ \beta_3 \cdot [A|B]_{ip} + \beta_4 \cdot [A \wedge B]_{ip} + \beta_5 \cdot [A \Rightarrow B]_{ip} \\ &+ \beta_6 \cdot [B \Rightarrow A]_{ip} + \beta_7 \cdot Other_{ip} + \epsilon_{ip} \end{aligned}$$

where p is a participant, i an item, *pos* is the item position (added with a quadratic term to model the overall speedup of responses), and $[\varphi]$ is coded 1 if the response is according to the prediction for φ, and 0 otherwise (the conditional event, $B|A$, interpretation is the baseline category thus does not appear as a predictor). The coefficient γ_{0p} represents between-participant variation in mean reaction time and γ_{1i} represents participant-invariant effects of items.

First the effect of the response type was tested. Adding this variable improved the fit of the model (ΔAIC $= -8$, log-likelihood ratio (LLR) $\chi^2(5) = 18.1$, $p = .003$). As predicted, conjunction responses were faster than the conditional event (95% CI $\in [-0.15, -0.04]$). The mean difference predicted using the model's fixed effect terms was 503 ms. Confidence intervals for all other

[4]Models were fitted using the lme4 package [1] in R (`www.r-project.org`). HPD intervals were estimated using MCMC draws from the posterior distributions. Log-likelihood ratio tests, and Akaike's information criterion (AIC), derived from the maximum log-likelihood estimates and penalized for the number of parameters, were used to compare fitted models.

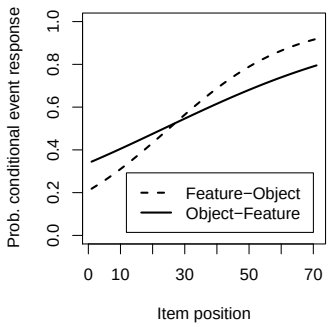


Figure 4: Interaction between item position and object-feature order.

strategy-type predictors versus the conditional event included 0. To the base model we also added main effects of sex and object-feature order, and two- and three-way interactions between these variables and the response type. There was a hint of an effect of the three-way interaction (LLR $\chi^2(5) = 10.1$, $p = .07$), but since $\Delta\text{AIC} = 0$ (suggesting over-fitting) and the effect is so weak, for the sake of model parsimony we removed it. Next we tested two-way interactions, in the presence of all others. There was an interaction between sex and the response type ($\Delta\text{AIC} = -6$, LLR $\chi^2(5) = 15.6$, $p = .008$), object-feature order and the response type ($\Delta\text{AIC} = -3$, LLR $\chi^2(5) = 13.2$, $p = .02$), but we found no evidence of an interaction between sex and object-feature order (LLR $\chi^2(1) = 0$). We simplified the model accordingly. The coefficient for the conjunction response versus the conditional event was still negative (95% CI $\in [-0.30, -0.13]$). There was a weak main effect that males were faster than females (95% CI $\in [-0.23, -0.01]$) and complicated interactions with the response type, interpretation of which we defer to another occasion. There was no main effect of object-feature order, but participants were slower when giving a conjunction response in the feature-object condition (95% CI $\in [0.06, 0.26]$).

We also investigated whether the tendency to interpret the material according to the conditional event was affected by the object-feature order. A generalized linear mixed effect model was fitted with binomial errors and a *logit* link. The dependent variable was the probability of a conditional event response. As before predictors were added for item position (*pos*). Also a predictor, *order*, was added for the object-feature order: 1 if feature-object and 0 if object-feature. We found no main effect of object-feature order ($\Delta\text{AIC} = 2$, LLR $\chi^2(1) = 0.03$, $p = .9$), however there was an interaction between item position and object-feature order ($\Delta\text{AIC} = -17$, LLR $\chi^2(1) = 18.5$, $p < .001$). The final model chosen was as follows:

$$logit(P(y_{ip} = 1)) = \beta_0 + \gamma_{0p} + \gamma_{1i} + \beta_1 \cdot pos_{ip} + \beta_2 \cdot pos_{ip} \cdot order_{ip} + \epsilon_{ip}$$

Figure 4 shows predictions from the model's fixed-effect estimates. At the beginning of the task, participants in the object-feature condition were more likely to use a conditional event interpretation than in the feature-object condition.

4 Discussion

The conditional event was the most common interpretation of the if-then (modal response for 76% of participants in Experiment 1 and 69% in Experiment 2), followed by the conjunction (12% in Experiment 1 and 24% in Experiment 2). Material conditional responses were rare. We provided evidence of interpretation shifts: 55% of participants shifted to a conditional event response during the task, and 80% of these shifted from conjunction responses. Conjunction is consistent with an implicit model in mental models theory [5], however the theory would predict a shift to the material conditional rather than to the conditional event. Changes of interpretation have been observed in an experiment using a non-probabilistic truth table task [12]: participants changed from a conjunction interpretation to either equivalence or the material conditional, and from equivalence to the material conditional. This effect was argued to be cued by the process of going through the truth table cases. We have also provided evidence that the shift to the conditional event interpretation was later for feature-object order compared to the object-feature order, extending a result from analogies processing [6] to an uncertain reasoning task.

It is difficult to distinguish between effects due to individual differences in interpretation and those due to differences in derivation. It seems unlikely, however, that a shift from a conjunction response to a conditional event response—the most common kind of shift—would be due to a change in derivation strategy. If people had a fixed interpretation of conditional event but got better at derivation, this would result in a shift from noise (giving an 'other' classification) to the conditional event. Only three participants shifted in this way. Therefore it is more likely that it is the interpretation that shifts and not the derivation.

Insight is often defined as the effect of suddenly understanding how to solve a problem after a period of impasse, often accompanied with an 'Aha!' feeling [2]. Our results suggest that participants who shifted interpretation demonstrated such an effect, both by qualitative shifts in response type, and also (for some participants) by spontaneous self-reports of insight. Problems used to study insight, e.g., anagrams, usually have a clear goal; the difficulty comes from how to achieve that goal from the starting state. For our reasoning task, however, the difficulty is in understanding what the goal is, i.e., what probability should be computed. The interpretation shift is thus a shift in understanding of the goal, rather than how to achieve the goal (simple counting). Another difference in our task is that there was no impasse: participants continue to do the best they can with their first interpretation. Again this is because they have a clear goal, however transitory, and know how to achieve it. Although the shift is sudden, it is still possible that parallel competing processes incrementally compute two (or many more) interpretations, then after some time, the most likely interpretation is inferred to be the conditional event. A similar incremental account has been given of sudden 'pop-out' solutions in anagram solving [7].

These results have important implications for building process models. Not only do different people reason to different interpretations, but individuals shift interpretations during a task. Studying trajectories of interpretation change reveals participants' inferences about correctness of interpretation. It is thus interesting that so many participants converge on the conditional event. Future work is needed to clarify when and for whom these shifts of interpretation occur, and what cues can facilitate or impede the process.

References

[1] D. Bates, M. Maechler, and B. Dai. *lme4: Linear mixed-effects models using S4 classes*, 2008. R package version 0.999375-28.

[2] E. Bowden, M. Jung-Beeman, J. Fleck, and J. Kounios. New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7):322–328, 2005.

[3] G. Coletti and R. Scozzafava. *Probabilistic Logic in a Coherent Setting*. Kluwer Academic Publishers, Dordrecht, 2002.

[4] J. St. B. T. Evans, S. J. Handley, and D. E. Over. Conditionals and conditional probability. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 29(2):321–335, 2003.

[5] P. N. Johnson-Laird and R. M. J. Byrne. Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109(4):646–677, 2002.

[6] G. D. Kleiter. Solving analogies by building propositions. In F. Klix and H. Hagendorf, editors, *Human Memory and Cognitive Capabilities*, pages 977–986. Elsevier, Amsterdam, 1986.

[7] L. R. Novick and S. J. Sherman. On the nature of insight solutions: Evidence from skill differences in anagram solution. *The Quarterly Journal of Experimental Psychology, A*, 56(2):351–382, 2003.

[8] M. Oaksford and N. Chater. *Bayesian Rationality: the probabilistic approach to human reasoning*. Oxford University Press, 2007.

[9] K. Oberauer and O. Wilhelm. The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4):680–693, 2003.

[10] N. Pfeifer and G. D. Kleiter. Coherence and nonmonotonicity in human reasoning. *Synthese*, 146(1–2):93–109, 2005.

[11] N. Pfeifer and G. D. Kleiter. Framing human inference by coherence based probability logic. *Journal of Applied Logic*, 7(2):206–217, 2009.

[12] G. Politzer. Differences in interpretation of implication. *The American Journal of Psychology*, 94(3):461–477, 1981.

[13] F. P. Ramsey. General propositions and causality. In D. H. Mellor, editor, *Philosophical Papers*, pages 145–163. Cambridge Unversity Press, 1929/1990. A scan of the original handwritten manuscript is available at `http://www.dspace.cam.ac.uk/handle/1810/194722`.

[14] L. J. Rips. *The Psychology of Proof: Deductive Reasoning in Human Thinking*. The MIT Press, Cambridge, MA, USA, 1994.

[15] K. Stenning and M. van Lambalgen. *Human reasoning and cognitive science*. MIT Press, Cambridge, Massachusetts, USA, 2008.

On an Approximative Solution to the Marginal Problem

Martin Janžura
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
janzura@utia.cas.cz

Abstract

With the aid of the Maximum Entropy principle, a solution to the marginal problem is obtained in a form of parametric exponential (Gibbs-Markov) distribution. The unknown parameters can be calculated by an optimization procedure that agrees with the maximum likelihood estimate but it is numerically hardly feasible for highly dimensional systems. A numerically easily feasible solution can be obtained by the algebraic Möbius formula. The formula, unfortunately, involves terms that are not directly available but can be approximated. And the main aim of the present paper consists in this approximation.

1 Introduction

We address the so-called marginal problem, i.e. the problem of reconstruction of a joint (global) distribution from a collection of marginal (local) ones. To the contrary with some other approaches, where the problem is studied either by graphical or combinatorial reasoning, or by iterative computational algorithms (see, e.g., [6] or [7]), here the solution is inspired, more-or-less, by a "statistical" point of view.

In order to find a unique representing joint distribution for the system, we employ the maximum entropy principle. Then, providing some technical assumptions being satisfied, the solution agrees with a parametric exponential (Gibbs) distribution as the most natural and convenient representative. The distribution is also Markovian with the neighborhood system induced by the system of marginals (Section 5.). Thus the structure of the distribution is known but the parameters are given only implicitly. In order to fix the parameters, we have to solve the same task as within the problem of statistical estimation. In particular, the parameters are obtained by an optimization procedure that agrees with the maximum likelihood (ML) estimate (as if the marginals were obtained from data). Thus, we may imagine the "input" information contained in the system of marginal (local) distributions as an evidence, and the problem of finding the unknown joint distribution is re-formulated as a parameter estimation problem. But, as it is well known, under a certain size of the model, any direct optimization method is unfeasible. Therefore, for calculating parameters of the representing distributions in full generality we need to apply some sim-

ulation procedure, usually based on the Markov Chain Monte Carlo methods (see Section 6.).

But, as we show finally in Section 7., we can also apply the combinatorial Möbius formula for direct evaluating the potentials of the Gibbs distributions, and these potentials are equal exactly to the unknown parameters.

Unfortunately, the formula involves marginals over larger sets of nodes, namely over the neighborhoods of particular nodes. Thus, for an easy calculation of the maximal entropy solution to the marginal problem , we have, at first, to extend the original marginals to these larger sets, at least approximately.

An approximative method of these extension, based partly on the ideas introduced in [6] or [7], is presented in Section 8. as the main goal of this paper.

For many topics of the present paper [7] or [9] are the basic references. For exponential distributions see [1]. For stochastic gradient method see [9] or [10], for general MCMC simulations see [2]. For the marginal problem see, e.g., [6] and the references therein. Some specific approaches can be found, e.g., in [4] or [8].

2 Basic definitions

Let us consider a finite set S of indices (sites, variables, nodes), and the space of configurations

$$\mathcal{X}_S = \bigotimes_{s\in S} \mathcal{X}_s$$

where $\mathcal{X}_s$ is a finite state space for every $s \in S$. For every $V \subset S$ we denote by $\mathrm{Pr}_V : \mathcal{X}_S \to \mathcal{X}_V$ the projection onto the space $\mathcal{X}_V = \bigotimes_{s\in V} \mathcal{X}_s$, and by $\mathcal{B}_V = \sigma(\mathrm{Pr}_V)$ the σ-algebra of cylinder (local) sets.

Further, by $\mathcal{P}_V$ we denote the class of all probability measures on $\mathcal{B}_V$, and by $\mathcal{F}_V$ the class of all real-valued $\mathcal{B}_V$-measurable functions. ($\mathcal{P}_V$ can be alternatively understood as the set of probability measures on $\mathcal{X}_V$, and $\mathcal{F}_V$ as the set of functions on $\mathcal{X}_V$. We shall not distinguish these two modes.) For $P_V \in \mathcal{P}_V$ and $W \subset V$ we shall denote by $P_{V/W} \in \mathcal{P}_W$ its projection into the space $\mathcal{P}_W$, i.e., the corresponding marginal distribution. (Whenever no confusion may occur, we shall write directly P_W.) On the other hand, by $P_{A|B}$ for $A, B \subset S, A \cap B = \emptyset$, we denote the corresponding conditional distribution.

3 Problem

Let us consider a system of (non-void) subsets $\mathcal{V} \subset \exp \mathrm{S}$, satisfying $V \setminus W \neq \emptyset$ for $V, W \in \mathcal{V}, V \neq W$, and a collection of *marginal distributions*

$$\mathcal{Q} = \{Q_V\}_{V\in\mathcal{V}}$$

where

$$Q_V \in \mathcal{P}_V \quad \text{for every } V \in \mathcal{V}.$$

Let us denote

$$\mathcal{P}_\mathcal{Q} = \{P_S \in \mathcal{P}_S; P_{S/V} = Q_V \quad \text{for every } V \in \mathcal{V}\}.$$

If $\mathcal{P}_\mathcal{Q} \neq \emptyset$ we quote the collection $\mathcal{Q}$ as *strongly consistent*.

The problem to be solved now consists in finding a suitable representative

$$\overline{P}_S \in \mathcal{P}_{\mathcal{Q}},$$

providing $\mathcal{Q}$ is strongly consistent.

4 Maximum entropy principle

Whenever $|\mathcal{P}_{\mathcal{Q}}| > 1$ we have to employ some additional criterion for selecting $\overline{P}_S$, which, in our case, will be the *maximum entropy principle* . For a justification of such approach see, e.g., [5] as the standard reference.

Let us recall the formulas for the *entropy* and the I *-divergence*, respectively, namely

$$H(P) = \int -\log P \,\mathrm{d}P = \sum_{x_S \in \mathcal{X}_S} -\log P(x_S)\, P(x_S),$$

and

$$I(P|Q) = \int \log \frac{P}{Q} \,\mathrm{d}P = \sum_{x_S \in \mathcal{X}_S} \log \frac{P(x_S)}{Q(x_S)}\, P(x_S)$$

providing the terms are well defined. Otherwise we set $I(P|Q) = \infty$.

Thus, applying the maximum entropy principle , we seek for

$$\overline{P}_S \in \mathrm{argmax}_{P_S \in \mathcal{P}_{\mathcal{Q}}} H(P_S)$$

or, more generally,

$$\overline{P}_S \in \mathrm{argmin}_{P_S \in \mathcal{P}_{\mathcal{Q}}} I(P_S|R_S)$$

where $R_S \in \mathcal{P}_S$ is some fixed reference probability measure.

For the sake of brevity, we shall deal directly with the first definition, which, after all, agrees with the latter one for uniform R_S.

5 Gibbs-Markov distributions

Further, we shall quote $P_S \in \mathcal{P}$ as the *Gibbs distribution* with the *potential* $U = \{U_A\}_{A \in \mathcal{A}}$ where $U_A \in \mathcal{F}_A$ for every $A \in \mathcal{A} \subset \exp \mathrm{S}$ (see, e.g., [9] for detailed treatment) if

$$P_S(y_S) \propto \exp\{\sum_{A \in \mathcal{A}} U_A(y_A)\}.$$

Then we shall write $P_S = P_S^U$. Moreover, since

$$P^U_{\{s\}|S\setminus\{s\}}(y_{\{s\}}|y_{S\setminus\{s\}}) \propto \exp\{\sum_{A \in \mathcal{A}, A \ni s} U_A(y_A)\},$$

P_S^U is also *Markovian* with the *neighborhood system* $\partial = \{\partial(s)\}_{s \in S}$ given by

$$t \in \partial(s) \quad \text{iff} \quad \{t, s\} \subset A \quad \text{for some} \quad A \in \mathcal{A}.$$

6 Maximum entropy solution

Now, let us fix a configuration $0_S \in \mathcal{X}_S$. For $B \subset S$ we denote $\mathcal{X}_B^0 = \bigotimes_{b\in B}(\mathcal{X}_b \setminus \{0_b\})$. Further, we denote

$$\overline{\mathcal{V}} = \{W \subset S;\, \emptyset \neq W \subset V \quad \text{for some } V \in \mathcal{V}\}.$$

Let as consider the class of potentials

$$\mathcal{U}^0 = \{U = (U_W)_{W\in\overline{\mathcal{V}}};\, U_W \in \mathcal{F}_W \text{ and } U_W(x_W) = 0 \text{ for every } x_W \in \mathcal{X}_W \setminus \mathcal{X}_W^0\}.$$

Then $\mathcal{U}^0$ is the space of so-called *vacuum potentials* (see, e.g., [3]). We may also write $U_W = \sum_{x_W\in\mathcal{X}_W^0} U_W(x_W)\delta_{x_W}$ with some real constants $\{U_W(x_W)\}_{x_W\in\mathcal{X}_W^0}$ for every $W \in \overline{\mathcal{V}}$.

Proposition 1. Let $P_S^{\overline{U}} \in \mathcal{P}_\mathcal{Q}$ for some $\overline{U} \in \mathcal{U}^0$. Then $\overline{U} \in \mathcal{U}^0$ is given uniquely and

$$P_S^{\overline{U}} = \overline{P}_S = \operatorname{argmax}_{P_S\in\mathcal{P}_\mathcal{Q}} H(P_S).$$

Proof. See Proposition 1 and 2 in [3]. □

Remark 1. We shall omit here the question of existence of $P_S^{\overline{U}} \in \mathcal{P}_\mathcal{Q}$ (see also, e.g., [3]). Here it will be simply assumed. Let us emphasize that such assumption involves also the condition

$$Q_V > 0 \quad \text{for every } V \in \mathcal{V}.$$

That will make the further calculations much easier since we do not have to take care about zeros. □

We shall rather discuss the problem of numerical feasibility. Namely, the unknown parameters $\{\overline{U}_W(x_W)\}_{x_W\in\mathcal{X}_W^0, W\in\overline{\mathcal{V}}}$ should be identified by the condition

$$P_S^{\overline{U}} \in \mathcal{P}_\mathcal{Q}$$

which means

$$P_W^{\overline{U}}(x_W) = Q_W(x_W) \quad \text{for every} \quad x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}$$

or, equivalently, by

$$\overline{U} = \arg\max_{U\in\mathcal{U}^0} \left[\sum_{W\in\overline{\mathcal{V}}}\sum_{x_W\in\mathcal{X}_W^0} U_W(x_W)Q_W(x_W) - \log\sum_{x_S\in\mathcal{X}_S}\exp\{\sum_{W\in\overline{\mathcal{V}}} U_W(x_W)\}\right].$$

Both methods contain terms that involve summing over the set $\mathcal{X}_S$ which is numerically hardly feasible for large S.

Hence, the *stochastic gradient method* (cf. [9], Section 15.4, or [10]) was introduced, based on substituting the "theoretical" terms by their simulated counterparts. The Markov Chain Monte Carlo (MCMC) – or some similar method – can be used for the *simulation* (cf., e. g., [2] for a survey).

Let us recall here that, in principle, the above method of identifying the parameters agrees with the statistical parameter estimation, namely the *maximum likelihood (ML)*, or, equivalently, the *minimum I-divergence* method. The only difference consists in the fact that within the statistical estimation the collection $\{Q_W(x_W)\}_{x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}}$ is given as an "evidence" obtained from observed data, in particular $Q_W(x_W) = \widehat{P}_{S/W}(x_W)$ for every $x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}$ where $\widehat{P}_S$ is the empirical distribution.

7 Möbius formula

Nevertheless, due to the problems as described above, we prefer much more straightforward method, given by *Möbius formula* (see, e.g., [9]), for identifying the parameters. Let us introduce the formula, which is rather general, in a form suitable for our purposes.

Proposition 2.
Let us denote $\Phi(x_S) = \log P^{\overline{U}}_S(x_S)$ with $\overline{U} \in \mathcal{U}^0$. Then

$$\overline{U}_W(x_W) = \sum_{B \subset W} (-1)^{|W \setminus B|} \left[\Phi(x_B, 0_{S \setminus B}) - \Phi(0_S)\right]$$

for every $x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}$.

Proof.
The relation can be verified by direct substitution. See, e.g.,[3] or [9]. □

Now, by elementary rearrangements, we obtain

$$\overline{U}_W(x_W) = \sum_{B \subset W} (-1)^{|W \setminus B|} \left[\log \frac{P^{\overline{U}}_S(x_B, 0_{S \setminus B})}{P^{\overline{U}}_S(0_S)}\right] =$$

$$= \sum_{B \subset W \setminus \{s\}} (-1)^{|W \setminus B|} \left[\log \frac{P^{\overline{U}}_S(x_B, 0_{S \setminus B})}{P^{\overline{U}}_S(x_{B \cup \{s\}}, 0_{S \setminus \{B \cup \{s\}\}})}\right] =$$

$$= \sum_{B \subset W \setminus \{s\}} (-1)^{|W \setminus B|} \left[\log \frac{P^{\overline{U}}_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s) \setminus \{B \cup \{s\}\}})}{P^{\overline{U}}_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s) \setminus \{B \cup \{s\}\}})}\right]$$

for every $x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}$, where $s \in W$ is arbitrary fixed. For the last expression note that $(W \setminus \{s\}) \subset \partial(s)$ since $s \in W$.

Now, suppose we are able to *extend* the original system of marginals $\mathcal{Q}$ **consistently** into the system

$$\mathcal{Q}^\partial = \{Q_{\overline{\partial}(s)}\}_{s \in S},$$

where $\overline{\partial}(s) = \partial(s) \cup \{s\}$, i.e.

$$P^{\overline{U}}_S \in \mathcal{P}_{\mathcal{Q}} \cap \mathcal{P}_{\mathcal{Q}^\partial}$$

can be guaranteed. Then we can calculate the parameters $\{\overline{U}_W(x_W)\}_{x_W\in\mathcal{X}^0_W, W\in\overline{\mathcal{V}}}$ directly from the Möbius formula, namely

$$\overline{U}_W(x_W) = \sum_{B\subset W\setminus\{s\}} (-1)^{|W\setminus B|}\left[\log\frac{Q_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}{Q_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}\right]$$

for every $x_W\in\mathcal{X}^0_W, W\in\overline{\mathcal{V}}$.

Actually, we do not need to know the complete distributions $Q_{\overline{\partial}(s)}, s\in S$, but only $Q_{\overline{\partial}(s)}(x_W, 0_{\overline{\partial}(s)\setminus W})$ for every $W\in\overline{\mathcal{V}}, W\subset\overline{\partial}(s)$, and $x_W\in\mathcal{X}_W$.

8 Approximation

Unfortunately, the exact extension is usually hardly available, but, from the practical point of view, a reasonable *approximation* can be sufficient. Let us continue with the above reasoning in order to obtain

$$\begin{aligned}\overline{U}_W(x_W) &= \sum_{B\subset W\setminus\{s\}} (-1)^{|W\setminus B|}\left[\log\frac{Q_{\overline{\partial}(s)}(x_B, 0_{\overline{\partial}(s)\setminus B\}})}{Q_{\overline{\partial}(s)}(x_{B\cup\{s\}}, 0_{\overline{\partial}(s)\setminus\{B\cup\{s\}\}})}\right]\\ &= \sum_{B\subset W} (-1)^{|W\setminus B|}\left[\log\frac{Q_{\overline{\partial}(s)}(x_B, 0_{\overline{\partial}(s)\setminus B\}})}{Q_{\overline{\partial}(s)}(0_{\overline{\partial}(s)})}\right].\end{aligned}$$

Now, for approximating $Q_{\overline{\partial}(s)}$ we shall use a rather standard *product form* (see, e.g., [4],[6], or [7]), but, first of all, we need some more notation.

For $s\in S$ we denote $\mathcal{V}_s = \{V\in\mathcal{V}; V\ni s\}$, $v_s = |V_s|$, and I_s the set of all possible enumerations of the elements of $\mathcal{V}_s$. Then, for every $\rho\in I_s$, we may set

$$\widehat{Q}^{\rho}_{\overline{\partial}(s)} = \frac{\prod_{A\in\mathcal{V}_s} Q_A}{\prod_{j=1,\ldots,v_s} Q_{B^\rho_j}}$$

where $B^\rho_j = A_{\rho(j)}\cap(\bigcup_{i=1}^{j-1} A_{\rho(i)})$ for every $j=1,\ldots,v_s$, as a natural estimate.

Remark 2. The product form for $\widehat{Q}^{\rho}_{\overline{\partial}(s)}$ can be also justified by the assumption

$$\frac{Q_{\overline{\partial}(s)}(x_B, 0_{\overline{\partial}(s)\setminus B\}})}{Q_{\overline{\partial}(s)}(x_{B\cup\{s\}}, 0_{\overline{\partial}(s)\setminus\{B\cup\{s\}\}})} = \frac{Q_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}{Q_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})} =$$

$$= \frac{P^{\overline{U}}_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}{P^{\overline{U}}_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}$$

where the latter term can be factorized by definition (see Section 5.). □

Further, for every pair $A, W\in\overline{\mathcal{V}}$ let us denote

$$\hat{u}_{W,A}(x_W) = \sum_{B\subset W} (-1)^{|W\setminus B|}\left[\log\frac{Q_A(x_{A\cap B}, 0_{A\setminus B\}})}{Q_A(0_A)}\right].$$

Proposition 3. Let $W \setminus A \neq \emptyset$. Then $\hat{u}_{W,A} \equiv 0$.

Proof. We may write

$$\hat{u}_{W,A}(x_W) = \sum_{B_1 \subset W \cap A} \sum_{B_2 \subset W \setminus A} (-1)^{|(W \cap A) \setminus B_1|} (-1)^{|(W \setminus A) \setminus B_2|} \left[\log \frac{Q_A(x_{A \cap B_1}, 0_{A \setminus B_1\}})}{Q_A(0_A)} \right].$$

And for $W \setminus A \neq \emptyset$ we have

$$\sum_{B_2 \subset W \setminus A} (-1)^{|(W \setminus A) \setminus B_2|} = 0.$$

□

With the above defined terms, by substituting the estimate $\widehat{Q}^{\rho}_{\overline{\partial}(s)}$ into the expression for $\overline{U}_W$, we may introduce the **approximation**

$$\widehat{U}^{s,\rho}_W = \sum_{A \in \mathcal{V}_s, A \supset W} u_{W,A} - \sum_{j=1,\ldots,v_s : B^{\rho}_j \supset W} u_{W,B^{\rho}_j}$$

for every $W \in \overline{\mathcal{V}}, s \in W$, and $\rho \in I_s$. Since $s \in W$ and $\rho \in I_s$ are "free parameters", we may, finally **average** over all possible choices, and obtain

$$\widehat{U}_W = |W|^{-1} \sum_{s \in W} |I_s|^{-1} \sum_{\rho \in I_s} \widehat{U}^{s,\rho}_W$$

for every $W \in \overline{\mathcal{V}}$.

Remark 3. It is apparent that for large W many terms disappear. E.g., for $W \in \mathcal{V}$ we have actually $\widehat{U}^{s,\rho}_W = u_{W,W}$ for every $s \in W$ and $\rho \in I_s$, and therefore also $\widehat{U}_W = u_{W,W}$. □

Remark 4. The main advantage of the method is given by its non-sensitivity to the break of assumptions. In practise, we do not have to check the consistency assumption for the original system $\mathcal{Q}$, and the possible zeros can be substituted by some small $\epsilon > 0$. In addition, the model does not require any interconnections between the approximated potential functions $\widehat{U}_W, W \in \overline{\mathcal{V}}$. □

Remark 5. With the model parameters $\{\overline{U}_W(x_W)\}_{x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}}$ we can easily calculate the relative values of the probability, namely $P^{\overline{U}}_S(x_S)/P^{\overline{U}}_S(y_S)$ for $x_S, y_S \in \mathcal{X}_S$, and the conditional distributions $P^{\overline{U}}_{A|\partial A}$ for "small" $A \subset S$. More complex terms can be again simulated, similarly as in Section 6. □

Acknowledgement.

Supported by grant GA ČR No.201/09/1931 and Research Center DAR (MŠMT ČR Project No. 1M0572).

References

[1] Barndorff–Nielsen, O. E. (1978) *Information and Exponential Families in Statistical Theory.* John Wiley and Sons, New York .

[2] Gilks W. R., Richardson S.,and Spiegelhalter D. J. (eds.) (1996) *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

[3] Janžura M. (2007) *On the connection between marginal problem, statistical estimation, and Möbius formula.* Kybernetika 43(5):619-631.

[4] Janžura M. and Boček P. (1998) *A method for knowledge integration.* Kybernetika 34 , 1, 41–55.

[5] Jaynes E. T. (1982) On the rationale of tmaximum entropy methods. *Proc. IEEE 70*, 939–952.

[6] Jiroušek R. and Vejnarová J. (2003) Construction of multidimensional model by operators of composition: current state of art. *Soft Computing 7*, 328–335.

[7] Lauritzen S. L.(1996) *Graphical Models.* University Press, Oxford .

[8] Perez A. and Studený. M.(2006)Comparison of two methods for approximation of probability distributions with prescribed marginals. *submitted to Kybernetika*

[9] Winkler G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer–Verlag, Berlin .

[10] Younes L. (1988) Estimation and annealing for Gibbsian fields. *Ann. Inst. Henri Poincare 24* , 2, 269–294.

There are Combinations and Compositions in Dempster-Shafer Theory of Evidence

Radim Jiroušek, Jiřina Vejnarová*
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
and
University of Economics, Prague
radim@utia.cas.cz, vejnar@utia.cas.cz

Abstract

It is a generally accepted fact that the Dempster's rule of combination plays a key role in Dempster-Shafer Theory of Evidence. In this paper the authors compare this combination rule with another one, which is called composition, and which was designed to create multidimensional basic assignments from a system of low-dimensional ones. The goal of this paper is to show that though the mentioned methods of combination were designed for totally different reasons, they manifest some similar formal properties and under very special conditions they even coincide.

1 Introduction

Dempster's rule of combination is often used as a method of fusion of several sources of information: combining two subjective evaluations of beliefs one can get a "summarized" evaluation expressing knowledge from both the considered sources (e.g. [6, 1, 4]).

It is not the goal of this paper to bring arguments for or against the above mentioned way of interpretation of the Dempster's rule of combination. Our goal is to compare this rule of combination with another combining tool, so called *operator of composition*, proposed for construction of multidimensional models from a number of low-dimensional ones. Here we do not consider fusion in its proper meaning. The purpose why the operator of composition was designed was not to fuse imprecise descriptions about the same object but to compose a number of descriptions each of them describing different properties of the object to get its global description. Using the terminology of AI, operator of composition was proposed to construct a model of global knowledge from

*The research was partially supported by Ministry of Education of the Czech Republic under grant no. 2C06019, and by Czech Science Foundation under the grants no. ICC/08/E010 and 201/09/1891.

a system of pieces of local knowledge. So, it corresponds to the process of *knowledge integration.*

Keeping this in mind, it is quite natural that we do not want to compare the mentioned two ways of combination to show that one of them is better than the other. Having been inspired by an anonymous referee of [3], we want to compare them from the formal point of view, because, though they were designed for different purposes, they manifest some similar properties, and they even coincide under some very special situations.

2 Notation and basic notions

2.1 Set notation

In the whole paper we will deal with a finite number of variables $X_1, X_2, \ldots, X_n$ each of which is specified by a finite set $\mathbf{X}_i$ of its values. So, we will consider multidimensional space of discernment

$$\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and its *subspaces*. For $K \subset N = \{1, 2, \ldots, n\}$, $\mathbf{X}_K$ denotes a Cartesian product of those $\mathbf{X}_i$, for which $i \in K$:

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

A *projection* of $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1, i_2, \ldots, i_\ell\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_\ell}) \in \mathbf{X}_K.$$

Analogously, for $K \subset L \subseteq N$ and $A \subset \mathbf{X}_L$, $A^{\downarrow K}$ will denote a *projection* of A into $\mathbf{X}_K$:

$$A^{\downarrow K} = \{y \in \mathbf{X}_K : \exists x \in A \;\; (y = x^{\downarrow K})\}.$$

Let us remark that we do not exclude situations when $K = \emptyset$. In this case $A^{\downarrow \emptyset} = \emptyset$.

In addition to the projection, in this text we will need also the opposite operation which will be called join. By a *join* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ we will understand a set

$$A \otimes B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \;\; \& \;\; x^{\downarrow L} \in B\}.$$

Notice that if K and L are disjoint then their join is just their Cartesian product

$$A \otimes B = A \times B.$$

If $K = L$ then

$$A \otimes B = A \cap B. \tag{1}$$

If $K \cap L \neq \emptyset$ and $A^{\downarrow K \cap L} \cap B^{\downarrow K \cap L} = \emptyset$ then also $A \otimes B = \emptyset$. Generally,

$$A \otimes B = (A \times \mathbf{X}_{L \setminus K}) \cap (B \times \mathbf{X}_{K \setminus L}). \tag{2}$$

2.2 Basic assignment notation

The role of a probability distribution from a probability theory is in Dempster-Shafer theory played by a basic (*probability or belief*) assignment. In this paper we shall use exclusively normalized basic assignments.

A *basic assignment* m on $\mathbf{X}_K$ is a function

$$m : \mathcal{P}(\mathbf{X}_K) \longrightarrow [0, 1],$$

for which $m(\emptyset) = 0$ and

$$\sum_{A \subseteq \mathbf{X}_K} m(A) = 1.$$

A basic assignment on $\mathbf{X}_K$ is called *vacuous* if $m(\mathbf{X}_K) = 1$, and it is called *simple* basic assignment *focused* on A (for $\emptyset \neq A \subset \mathbf{X}_K$) if $m(A) = a$ for $a > 0$ and $m(\mathbf{X}_K) = 1 - a$.

If $m(A) > 0$, then A is said to be a *focal element* of m. If all the focal elements of m are singletons (i.e. $m(A) > 0$ implies that $|A| = 1$) then we say that m is *Bayesian.*

For $L \subset K$ and basic assignment m on $\mathbf{X}_K$ one gets its *marginal basic assignment* $m^{\downarrow L}$ by computing for each $B \subseteq \mathbf{X}_L$:

$$m^{\downarrow L}(B) = \sum_{A \subseteq \mathbf{X}_K : A^{\downarrow L} = B} m(A).$$

Conversely, let m be a basic assignment on $\mathbf{X}_L$. Its *vacuous extension* on $\mathbf{X}_K$ is defined for all $A \subseteq \mathbf{X}_K$ in the following way

$$m^{\uparrow K}(A) = \begin{cases} m(A^{\downarrow L}) & \text{if } A = A^{\downarrow L} \times \mathbf{X}_{K \setminus L}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

2.3 Dempster's rule of combination

Dempster's rule of combination is usually defined for two basic assignments m_1, m_2 defined on the same frame of discernment (say $\mathbf{X}_K$) by the formula

$$(m_1 \oplus m_2)(C) = \frac{\sum\limits_{A,B \subseteq \mathbf{X}_K A \cap B = C} m_1(A) m_2(B)}{1 - \sum\limits_{A,B \subseteq \mathbf{X}_K : A \cap B = \emptyset} m_1(A) m_2(B)}, \tag{4}$$

for each $C \subseteq \mathbf{X}_K$. For the purpose of this paper we need its generalization to cover situations when one wants to combine two basic assignments, which are not defined on the same frame of discernment. Regarding equality (1), the natural generalization, which will be used in this paper, is the one introduced in the following definition.

Definition 1. For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ ($K \neq \emptyset \neq L$) their *combination* is computed according to the formula (for all

$C \subseteq \mathbf{X}_{K\cup L}$)[1]:

$$(m_1 \oplus m_2)(C) = \frac{\sum\limits_{A\subseteq\mathbf{X}_K}\ \sum\limits_{B\subseteq\mathbf{X}_L : A\otimes B=C} m_1(A)m_2(B)}{1 - \sum\limits_{A\subseteq\mathbf{X}_K}\ \sum\limits_{B\subseteq\mathbf{X}_L : A\otimes B=\emptyset} m_1(A)m_2(B)}.$$

Substituting vacuous extensions of m_1 and m_2 on $\mathbf{X}_{K\cup L}$ into formula (4), one gets

$$\begin{aligned}(m_1 \oplus m_2)(C) &= \frac{\sum\limits_{A,B\subseteq\mathbf{X}_{K\cup L} A\cap B=C} m_1^{\uparrow K\cup L}(A)m_2^{\uparrow K\cup L}(B)}{1 - \sum\limits_{A,B\subseteq\mathbf{X}_{K\cup L} : A\cap B=\emptyset} m_1^{\uparrow K\cup L}(A)m_2^{\uparrow K\cup L}(B)} \\ &= \frac{\sum\limits_{D\subseteq\mathbf{X}_K}\ \sum\limits_{E\subseteq\mathbf{X}_L : (D\times\mathbf{X}_{L\setminus K})\cap(E\times\mathbf{X}_{K\setminus L})=C} m_1(D)m_2(E)}{1 - \sum\limits_{D\subseteq\mathbf{X}_K}\ \sum\limits_{E\subseteq\mathbf{X}_L : (D\times\mathbf{X}_{L\setminus K})\cap(E\times\mathbf{X}_{K\setminus L})=\emptyset} m_1(D)m_2(E)},\end{aligned}$$

which is equivalent (taking into account expression (2)) the formula in Definition 1.

It is well known [5] that the following basic properties hold true for Dempster's rule of combination.

Lemma 1. *Let $K, L, M \subseteq N$. For arbitrary basic assignments m_1, m_2, m_3 defined on $\mathbf{X}_K, \mathbf{X}_L\mathbf{X}_M$, respectively:*

(i) *$m_1 \oplus m_2$ is a basic assignment on $\mathbf{X}_{K\cup L}$;*

(ii) *$m_1 \oplus m_2 = m_2 \oplus m_1$;*

(iii) *$(m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3)$.*

2.4 Operator of composition

An operator of composition was for basic assignments defined in [2] by the following definition.

Definition 2. For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ ($K \neq \emptyset \neq L$) a *composition* $m_1 \triangleright m_2$ is defined for each $C \subseteq \mathbf{X}_{K\cup L}$ by one of the following expressions:

[a] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$ and $C = C^{\downarrow K} \otimes C^{\downarrow L}$ then

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})};$$

[1] For the purpose of this paper we do not consider situations when

$$\sum_{A\subseteq\mathbf{X}_K}\ \sum_{B\subseteq\mathbf{X}_L : A\otimes B=\emptyset} m_1(A)m_2(B) = 1.$$

[b] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ and $C = C^{\downarrow K} \times \mathbf{X}_{L\setminus K}$ then

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

[c] in all other cases $(m_1 \triangleright m_2)(C) = 0$.

Example 1. Let $\mathbf{X}_1 = \{a, \bar{a}\}, \mathbf{X}_2 = \{b, \bar{b}\}$ and $\mathbf{X}_3 = \{c, \bar{c}\}$ be three frames of discernment and let us consider the following two simple basic assignments m_1 and m_2 defined on $\mathbf{X}_1 \times \mathbf{X}_2$ and $\mathbf{X}_2 \times \mathbf{X}_3$, respectively:

$$\begin{aligned} m_1(\mathbf{X}_1 \times \{b\}) &= 0.4, \\ m_1(\mathbf{X}_1 \times \mathbf{X}_2) &= 0.6, \\ m_2(\mathbf{X}_2 \times \{c\}) &= 0.5, \\ m_2(\mathbf{X}_2 \times \mathbf{X}_3) &= 0.5. \end{aligned}$$

From Definition 2 one can immediately see that the formula in case **[a]** can assign a positive value to $(m_1 \triangleright m_2)(A)$ and/or $(m_2 \triangleright m_1)(A)$ only for those $A \subseteq \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ for which

$$A^{\downarrow\{1,2\}} = \mathbf{X}_1 \times \{b\} \quad \text{or} \quad A^{\downarrow\{1,2\}} = \mathbf{X}_1 \times \mathbf{X}_2,$$

and

$$A^{\downarrow\{2.3\}} = \mathbf{X}_2 \times \{c\} \quad \text{or} \quad A^{\downarrow\{2,3\}} = \mathbf{X}_2 \times \mathbf{X}_3.$$

There are only two such sets, namely:

$$\mathbf{X}_1 \times \mathbf{X}_2 \times \{c\} \quad \text{and} \quad \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3.$$

For these sets we get

$$(m_1 \triangleright m_2)(\mathbf{X}_1 \times \mathbf{X}_2 \times \{c\}) = \frac{m_1(\mathbf{X}_1 \times \mathbf{X}_2) \cdot m_2(\mathbf{X}_2 \times \{c\})}{m_2^{\downarrow\{2\}}(\mathbf{X}_2)} = \frac{0.6 \cdot 0.5}{1} = 0.3,$$

$$(m_1 \triangleright m_2)(\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3) = \frac{m_1(\mathbf{X}_1 \times \mathbf{X}_2) \cdot m_2(\mathbf{X}_2 \times \mathbf{X}_3)}{m_2^{\downarrow\{2\}}(\mathbf{X}_2)} = \frac{0.6 \cdot 0.5}{1} = 0.3$$

and similarly

$$(m_2 \triangleright m_1)(\mathbf{X}_1 \times \mathbf{X}_2 \times \{c\}) = \frac{m_2(\mathbf{X}_2 \times \{c\}) \cdot m_1(\mathbf{X}_1 \times \mathbf{X}_2)}{m_1^{\downarrow\{2\}}(\mathbf{X}_2)} = \frac{0.5 \cdot 0.6}{0.6} = 0.5,$$

$$(m_2 \triangleright m_1)(\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3) = \frac{m_2(\mathbf{X}_2 \times \mathbf{X}_3) \cdot m_1(\mathbf{X}_1 \times \mathbf{X}_2)}{m_1^{\downarrow\{2\}}(\mathbf{X}_2)} = \frac{0.5 \cdot 0.6}{0.6} = 0.5.$$

Since $m_2(\{b\}) = 0$, from case **[b]** of Definition 2 we will get yet another focal element for $m_1 \triangleright m_2$, namely

$$A = \mathbf{X}_1 \times \{b\} \times \mathbf{X}_3,$$

for which

$$A^{\downarrow\{1,2\}} = \mathbf{X}_1 \times \{b\} \quad \text{and} \quad A^{\downarrow\{3\}} = \mathbf{X}_3.$$

Table 1: Composed basic assignments.

A	$(m_1 \triangleright m_2)(A)$	$(m_2 \triangleright m_1)(A)$
$\mathbf{X}_1 \times \mathbf{X}_2 \times \{c\}$	0.3	0.5
$\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$	0.3	0.5
$\mathbf{X}_1 \times \{b\} \times \mathbf{X}_3$	0.4	0

For this set we get

$$(m_1 \triangleright m_2)(\mathbf{X}_1 \times \{b\} \times \mathbf{X}_3) = m_1(\mathbf{X}_1 \times \{b\}) = 0.4.$$

Notice that when computing a composition $m_2 \triangleright m_1$, case **[b]** of Definition 2 does not assign a positive value to any subset A of $\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$, since if $m_2^{\downarrow\{2\}}(A^{\downarrow\{2\}}) > 0$ then also $m_1^{\downarrow\{2\}}(A^{\downarrow\{2\}}) > 0$.

Both the composed basic assignments $m_1 \triangleright m_2$ and $m_2 \triangleright m_1$ are outlined in Table 1 (recall once more that for all other $A \subseteq \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ different from those included in Table 1, both assignments equal 0). It is also evident from the table that the operator $\triangleright$ is not commutative.

Let us present the most important properties of the operator of composition for basic assignments, which were proved in [2].

Lemma 2. *Let $K, L \subseteq N$. For arbitrary basic assignments m_1, m_2 defined on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively:*

(i) *$m_1 \triangleright m_2$ is a basic assignment on $\mathbf{X}_{K \cup L}$;*

(ii) $m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K_1 \cap K_2} = m_2^{\downarrow K_1 \cap K_2}$;

(iii) $(m_1 \triangleright m_2)^{\downarrow K_1} = m_1$.

3 Relation of combinations and compositions

3.1 Disjoint domains

Theorem 1. *Let $K, L \subseteq N$ and m_1, m_2 be basic assignments defined on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. If $K \cap L = \emptyset$ then*

$$m_1 \triangleright m_2 = m_2 \triangleright m_1 = m_1 \oplus m_2.$$

Proof. For disjoint K, L and $A \subseteq \mathbf{X}_K, B \subseteq \mathbf{X}_L$ one gets $A \otimes B = A \times B$ and $m_2^{\downarrow K \cap L} \equiv 1$. Therefore, for computation of $m_1 \triangleright m_2$ (for any focal element $C \subseteq \mathbf{X}_{K \cup L}$ of $m_1 \triangleright m_2$) only case [a] of Definition 2 is employed, and therefore

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L}) = \sum_{A = C^{\downarrow K}} \sum_{B = C^{\downarrow L}} m_1(A) m_2(B)$$

$$= \sum_{A \subseteq \mathbf{X}_K} \sum_{B \subseteq \mathbf{X}_L : A \otimes B = C} m_1(A) m_2(B) = (m_1 \oplus m_2)(C),$$

because, in this case,

$$\sum_{A\subseteq\mathbf{X}_K}\ \sum_{B\subseteq\mathbf{X}_L:A\cap B=\emptyset} m_1(A)m_2(B)=0.$$

The fact that $m_1\triangleright m_2 = m_2\triangleright m_1$ follows immediately from property (ii) of Lemma 2. □

3.2 Identical domains

Theorem 2. *If for arbitrary two basic assignments m_1, m_2 on $\mathbf{X}_K$ each focal element of m_2 contains all the focal elements of m_1, i.e.*

$$m_1(A)>0, m_2(B)>0 \implies A\subseteq B,$$

then

$$m_1\triangleright m_2 = m_1\oplus m_2.$$

Proof. First, compute

$$\begin{aligned}\sum_{A,B\subseteq\mathbf{X}_K:A\otimes B=\emptyset} m_1(A)m_2(B) &= \sum_{A,B\subseteq\mathbf{X}_K:A\cap B=\emptyset} m_1(A)m_2(B)\\ &= \sum_{A\subseteq\mathbf{X}_K} m_1(A)\sum_{B\subseteq\mathbf{X}_K:A\cap B=\emptyset} m_2(B)=0,\end{aligned}$$

because, under the given assumptions, for each focal element A of m_1

$$\sum_{B\subseteq\mathbf{X}_K:A\cap B=\emptyset} m_2(B)=0.$$

Now, we can easily compute $(m_1\oplus m_2)(C)$ for any focal element C of m_1.

$$(m_1\oplus m_2)(C)=\sum_{A\subseteq\mathbf{X}_K} m_1(A)\sum_{B\subseteq\mathbf{X}_K:A\otimes B=C} m_2(B)=\sum_{A\subseteq\mathbf{X}_K:A=C} m_1(A)$$

$$= m_1(C).$$

In this way we obtained that $(m_1\oplus m_2)(C)=m_1(C)$ for all focal elements C of m_1. Therefore, since

$$\sum_{C\subseteq\mathbf{X}_K}(m_1\oplus m_2)(C)=\sum_{C\subseteq\mathbf{X}_K} m_1(C)=1,$$

it is clear that $(m_1\oplus m_2)(C)=m_1(C)$ for all $C\subseteq\mathbf{X}_K$, and therefore also

$$m_1\oplus m_2 = m_1 = m_1\triangleright m_2.$$ □

As a special case of Theorem 2 one gets the following assertion.

Corollary 1. *Let m_1 be an arbitrary basic assignment on $\mathbf{X}_K$ and let $\mathcal{F}$ denote the set of its focal elements. If m_2 is a simple basic assignment on $\mathbf{X}_K$ focused on B such that $B\supseteq\cup_{A\in\mathcal{F}}A$, then*

$$m_1\triangleright m_2 = m_1\oplus m_2.$$

3.3 General situation

Let us start studying general overlapping (but not identical) frames of discernment by an example illustrating the fact that a sufficient condition describing situations when combination and composition results in the same basic assignments cannot be obtained as a generalization of results from the previous two subsections.

Example 2. Consider two basic assignments m_1 on $\mathbf{X}_{\{1,2,3\}}$ and m_2 on $\mathbf{X}_{\{2,3,4\}}$ (with $\mathbf{X}_1 = \{a, \bar{a}\}, \mathbf{X}_2 = \{b, \bar{b}\}, \mathbf{X}_3 = \{c, \bar{c}\}, \mathbf{X}_4 = \{d, \bar{d}\}$), each having only two focal elements:

$$\begin{array}{lll} m_1: & A_1 = \{abc\}, A_2 = \{abc, \bar{a}\bar{b}\bar{c}\} & m_1(A_1) = 1/4, m_1(A_2) = 3/4. \\ m_2: & B_1 = \{bcd, \bar{b}\bar{c}d\}, B_2 = \{bcd, b\bar{c}d, \bar{b}\bar{c}\bar{d}\} & m_2(B_1) = 1/3, m_2(B_2) = 2/3. \end{array}$$

The reader can immediately see that each focal element of $m_2^{\downarrow\{2,3\}}$ contains all the focal elements of $m_1^{\downarrow\{2,3\}}$; i.e. $A_1^{\downarrow\{2,3\}} = \{bc\}$ and $A_2^{\downarrow\{2,3\}} = \{bc, \bar{b}\bar{c}\}$ are subsets of both $B_1^{\downarrow\{2,3\}} = \{bc, \bar{b}\bar{c}\}$ and $B_2^{\downarrow\{2,3\}} = \{bc, b\bar{c}, \bar{b}\bar{c}\}$.

Realizing that

$$\begin{aligned} A_1 \otimes B_1 &= \{abcd\}, \\ A_1 \otimes B_2 &= \{abcd\}, \\ A_2 \otimes B_1 &= \{abcd, \bar{a}\bar{b}\bar{c}d\}, \\ A_2 \otimes B_2 &= \{abcd, \bar{a}\bar{b}\bar{c}\bar{d}\}, \end{aligned}$$

it is clear that

$$\sum_{A \subseteq \mathbf{X}_{\{1,2,3\}}} \sum_{B \subseteq \mathbf{X}_{\{2,3,4\}}: A \otimes B = \emptyset} m_1(A) m_2(B) = 0,$$

and therefore

$$\begin{aligned} (m_1 \oplus m_2)(\{abcd\}) &= \sum_{A \subseteq \mathbf{X}_{\{1,2,3\}}} \sum_{B \subseteq \mathbf{X}_{\{2,3,4\}}: A \otimes B = \{abcd\}} m_1(A) m_2(B) \\ &= m_1(A_1) m_2(B_1) + m_1(A_1) m_2(B_2) = 1/4. \end{aligned}$$

When computing $m_1 \triangleright m_2$ one has to realize that even though

$$\{abcd\} = \{abcd\}^{\downarrow\{1,2,3\}} \otimes \{abcd\}^{\downarrow\{2,3,4\}},$$

$m_2^{\downarrow\{2,3\}}(\{bc\}) = 0$ and therefore neither case [a] nor [b] of Definition 2 is applicable for computing $(m_1 \triangleright m_2)(\{abcd\})$, and therefore it equals 0 according to case [c]. So we obtained that in this example $m_1 \oplus m_2 \neq m_1 \triangleright m_2$.

Theorem 3. *Let m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ be such basic assignments that each focal element A of m_1 and each focal element B of m_2 projects to a unique set in $\mathbf{X}_{K \cap L}$. Then*

$$m_1 \triangleright m_2 = m_1 \oplus m_2.$$

Proof. First, let us note that the assumption that all focal elements of both m_1 and m_2 project to a unique set implies, that $m_2^{\downarrow K \cap L}(A^{\downarrow K \cap L}) = 1$ for any focal element A of m_1.

Now, consider any $C \subseteq \mathbf{X}_{K\cup L}$ for which $C = C^{\downarrow K} \otimes C^{\downarrow L}$. For this C

$$(m_1 \oplus m_2)(C) = \frac{\sum_{A\subseteq\mathbf{X}_K} \sum_{B\subseteq\mathbf{X}_L: A\otimes B=C} m_1(A)m_2(B)}{1 - \sum_{A\subseteq\mathbf{X}_K} \sum_{B\subseteq\mathbf{X}_L: A\otimes B=\emptyset} m_1(A)m_2(B)}$$
$$\geq \sum_{A\subseteq\mathbf{X}_K} \sum_{B\subseteq\mathbf{X}_L: A\otimes B=C} m_1(A)m_2(B) \geq m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L}).$$

Simultaneously, if $m_1(C^{\downarrow K}) > 0$,

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})} = m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L}).$$

Since if $m_1(C^{\downarrow K}) = 0$ then also

$$(m_1 \triangleright m_2)(C) = 0 = m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L}),$$

one can see that for all $C \subseteq \mathbf{X}_{K\cup L}$ for which $C = C^{\downarrow K} \otimes C^{\downarrow L}$

$$(m_1 \oplus m_2)(C) \geq m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L}) = (m_1 \triangleright m_2)(C).$$

Regarding Definition 2, according to which $(m_1 \triangleright m_2)(C) = 0$ for $C \neq C^{\downarrow K} \otimes C^{\downarrow L}$, we see that

$$(m_1 \oplus m_2)(C) \geq (m_1 \triangleright m_2)(C)$$

holds true for all $C \subseteq \mathbf{X}_{K\cup L}$, from which, because both $m_1 \oplus m_2$ and $m_1 \triangleright m_2$ are normalized basic assignments, we get that $m_1 \oplus m_2 = m_1 \triangleright m_2$. □

Example 3. Let X_1, X_2 and X_3 be three binary variables with values in $\mathbf{X}_1 = \{a, \bar{a}\}, \mathbf{X}_2 = \{b, \bar{b}\}$, $\mathbf{X}_3 = \{c, \bar{c}\}$ and m_1 and m_2 be two basic assignments on $\mathbf{X}_1 \times \mathbf{X}_3$ and $\mathbf{X}_2 \times \mathbf{X}_3$ respectively, both of them having only two focal elements:

$$\begin{array}{lll} m_1: & A_1 = \{a\bar{c}, \bar{a}\bar{c}\}, A_2 = \{a\bar{c}, \bar{a}c\} & m_1(A_1) = 1/2, m_1(A_2) = 1/2. \\ m_2: & B_1 = \{b\bar{c}, \bar{b}\bar{c}\}, B_2 = \{b\bar{c}, \bar{b}c\} & m_2(B_1) = 1/2, m_2(B_2) = 1/2. \end{array} \quad (5)$$

One can immediately see that both $A_1 \otimes B_2$ and $A_2 \otimes B_1$ are empty and therefore $m_1 \oplus m_2$ has only two focal elements, namely $A_1 \otimes B_1 = \mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{c}\}$ and $A_2 \otimes B_2 = \{ab\bar{c}, \bar{a}\bar{b}c\}$. For these focal elements we have

$$(m_1 \oplus m_2)(\mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{c}\}) = \frac{m_1(A_1)m_2(B_1)}{1 - (m_1(A_1)m_2(B_2) + m_1(A_2)m_2(B_1))} = 1/2,$$
$$(m_1 \oplus m_2)(\{ab\bar{c}, \bar{a}\bar{b}c\}) = \frac{m_1(A_2)m_2(B_2)}{1 - (m_1(A_1)m_2(B_2) + m_1(A_2)m_2(B_1))} = 1/2$$

and simultaneously

$$(m_1 \triangleright m_2)(\mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{c}\}) = 1/2,$$
$$(m_1 \triangleright m_2)(\{ab\bar{c}, \bar{a}\bar{b}c\}) = 1/2.$$

Thus we got that for the basic assignments defined in expressions (5) $m_1 \oplus m_2 = m_1 \triangleright m_2$. Nevertheless, it does not mean that for any couple of basic assignments m_1, m_2 defined on $\mathbf{X}_1 \times \mathbf{X}_2$, $\mathbf{X}_2 \times \mathbf{X}_3$, respectively, with the

respective focal elements A_1, A_2 and B_1, B_2, the coincidence must hold. This happened because we chose special values of the considered basic assignments. If we change the values of m_1 and m_2 e.g. in the following way:

$$\begin{aligned} m_1'(A_1) = 1/3 \quad m_1'(A_2) = 2/3, \\ m_2'(B_1) = 1/3 \quad m_2'(B_2) = 2/3, \end{aligned}$$

we will get, analogously to (6),

$$\begin{aligned} (m_1' \oplus m_2')(\mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{c}\}) = 1/5, \\ (m_1 \oplus m_2)(\{ab\bar{c}, \bar{a}\bar{b}c\}) = 4/5, \end{aligned}$$

and

$$\begin{aligned} (m_1 \triangleright m_2)(\mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{c}\}) = 1/3, \\ (m_1 \triangleright m_2)(\{ab\bar{c}, \bar{a}\bar{b}c\}) = 2/3. \end{aligned}$$

Special property holds for Bayesian basic assignments.

Theorem 4. *Let $K, L \subseteq N$ and m_1, m_2 be Bayesian basic assignments defined on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. Then*

$$m_1 \triangleright m_2 = m_1 \oplus m_2$$

if $m_2^{\downarrow K \cap L}$ corresponds to uniform probability distribution.

Proof. The assumption that $m_2^{\downarrow K \cap L}$, being Bayesian basic assignment, corresponds to the uniform probability distribution implies that $m_2^{\downarrow K \cap L}$ is positive for any singleton from $\mathbf{X}_{K \cap L}$. This shows that case [b] of Definition 2 is not applicable to any $C \subseteq \mathbf{X}_{K \cup L}$ such that $C^{\downarrow K \cap L}$ is singleton.

Now consider an arbitrary singleton $C \subset \mathbf{X}_{K \cup L}$. It is obvious that $C = C^{\downarrow K} \otimes C^{\downarrow L}$ and therefore, according to case [a] of Definition 2,

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{\beta}, \tag{6}$$

where $\beta = m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L})$ is, due to the assumption posed on $m_2^{\downarrow K \cap L}$, the same for all singletons $C \subset \mathbf{X}_{K \cup L}$. On the other hand, if $C \subset \mathbf{X}_{K \cup L}$ is not singleton then either $C^{\downarrow K}$ or $C^{\downarrow L}$ cannot be singleton and therefore, if $(m_1 \triangleright m_2)(C)$ is assigned by case [a] of Definition 2, the value of $(m_1 \triangleright m_2)(C)$ is 0. In case that $(m_1 \triangleright m_2)(C)$ is assigned by case [b] of Definition 2, the resulting value is also 0, because this case is applicable only when $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L}) = 0$, which may appear only when $C^{\downarrow K \cap L}$ is not singleton and therefore neither $C^{\downarrow K}$ is a singleton, which means that $m_1(C^{\downarrow K}) = 0$. So, we showed that $m_1 \triangleright m_2$ is defined by (6) for singletons and for non-singletons it equals 0.

Let us denote

$$\alpha = \sum_{A \subseteq \mathbf{X}_K} \sum_{B \subseteq \mathbf{X}_L : A \otimes B = \emptyset} m_1(A) \cdot m_2(B).$$

For the considered Bayesian assignments

$$m_1(A) \cdot m_2(B)$$

can be positive only when both $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ are singletons. Therefore for any singleton $C \subseteq \mathbf{X}_{K\cup L}$

$$(m_1 \oplus m_2)(C) = \frac{\sum\limits_{A\subseteq\mathbf{X}_K} \sum\limits_{B\subseteq\mathbf{X}_L : A\otimes B=C} m_1(A)\cdot m_2(B)}{1-\alpha}$$

$$= \frac{m_1(C^{\downarrow K})\cdot m_2(C^{\downarrow L})}{1-\alpha}, \tag{7}$$

and for non-singletons C

$$(m_1 \oplus m_2)(C) = 0 = (m_1 \triangleright m_2)(C).$$

To prove the required equality

$$(m_1 \oplus m_2)(C) = (m_1 \triangleright m_2)(C)$$

also for singletons it is enough to compare equalities (7) and (6) and again realize that both $m_1 \oplus m_2$ and $m_1 \triangleright m_2$ are normalized basic assignments and therefore $1-\alpha = \beta$. □

4 Conclusions

In the paper we introduced the operator of composition for basic assignments and compared it with the famous Dempster's rule of combination. We showed that though Dempster's rule of combination and operator of composition were designed for different purposes they coincide in special situations; $m_1 \oplus m_2 = m_1 \triangleright m_2$

- when the combined basic assignments m_1 and m_2 are defined on disjoint frames of discernment;
- when all the focal elements of m_1 are contained in each focal element of m_2 and the basic assignments in question are defined on the same frame of discernment;
- when all the focal elements of both m_1 and m_2 project to the same subset of the overlapping frame of discernment.

Naturally, as shown in Example 3, the above described situations do not form a complete list of conditions under which the studied two operators coincide.

References

[1] Th. Denoeux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artificial Intelligence*, 172 (2008), pp. 234–264.

[2] R. Jiroušek, J. Vejnarová and M. Daniel. Compositional Models of Belief Functions. In: *Proc. of the 5th Int. Symposium on Imprecise Probabilitis and Their Applications ISIPTA'07*, (G. de Cooman, J. Vejnarová, M. Zaffalon, eds.). Mat-fyz Press, Praha, pp. 243-252, 2007.

[3] R. Jiroušek. On a Conditional Irrelevance Relation for Belief Functions Based on the Operator of Composition, (Gabriele Kern-Isberner, Ghristoph Beierle, eds.). To appear in: *Proc. of the KI07-Workshop on Dynamics of Knowledge and Belief*. Osnabrueck, Germany, 2007.

[4] A. Kallel, S. Hégarat-Mascle. Combination of partially non-distinct beliefs: The cautious-adaptive rule. *Int. J. Approx. Reason* (2009), doi:10.1016/j.ijar.2009.03.006.

[5] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey, 1976.

[6] Ph. Smets, Analyzing the combination of conflicting belief functions. *Information Fusion* **8** (4), 2007, pp. 387–412.

Structuring Essential Graphs *

Gernot D. Kleiter
Department of Psychology
University of Salzburg, Austria
gernot.kleiter@sbg.ac.at

Abstract

An essential graphs represents a class of Markov equivalent Bayesian networks. It may contain both directed and undirected edges. The contribution describes principles by which essential graphs can be enumerated in a hierarchical way.

We first generalize the concept of a terminal nodes so that it is applicable to graphs with directed and undirected edges. We define a symmetry relation for essential graphs. The number of labelings of an essential graph is easily obtained with the help of symmetries. We represent essential graphs as a special layered graphs admitting within-layer edges. Each layering determines a unique minimal backbone pattern of the edges that link neighboring layers. Within-layer edges build cliques of completely connected components. Long-distance edges are selected by symmetry properties. Possible applications to model learning and the assessment of prior distributions for model structures are discussed.

1 Introduction

We investigate the model space of essential graphs. Knowledge about the space of possible models is a prerequisite for introducing prior probabilities in model learning. Knowledge about different subclasses of models can be of considerable help for various search strategies in model learning. Our long-term goal is a hierarchical subdivision of the class of models. In the future such a hierarchy can be used to assess and analyse the probability of whole graphical structures in model learning.

Essential graphs were characterised and their properties extensively analysed by Andersson, Madigan, and Perlman [1]. Essential graphs are closely related to Bayesian networks. Why are we working with essential graphs and not with Bayesian networks? For counting and enumeration Bayesian networks are inappropriate. The structure of any conditional independence model is characterised by a set of conditional independences. If this set is represented by a directed acyclic graph, i.e., by a Baysian network, then the representation is not necessarily unique. One and the same set of conditional independences may correspond to different Bayesian networks. Thus Bayesian networks lead to multiple counts.

*Supported by the Austrian Science Fund, I141, within the European Science Foundation EUROCORES programme LogICCC.

This property is well known and called *Markov equivalence*. The three Bayesian networks (a), (b), and (c) in Figure 1 represent the conditional independence of B and C given A. They are Markov equivalent. A vee-structure in panel (e) is both, a Bayesian network and an essential graph. It encodes uniquely the marginal independence of A and B.

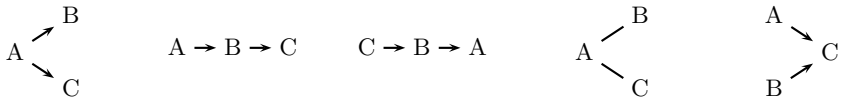


Figure 1: (a), (b), and (c) show three Markov equivalent Bayesian networks and (d) their unique representation as an essential graph. (e) shows a unique structure which is both a Bayesian network and an essential graph.

Typically essential graphs are hybrid and consists of both directed and undirected edges. But there are also models in which all edges are undirected and cases in which all edges are directed. Figure 2 shows another example of Markov equivalent Bayesian networks, namely complete graphs. They are all represented by the undirected essential graph (g).

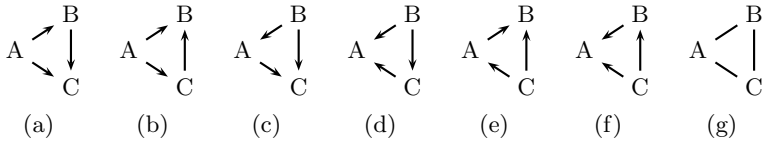


Figure 2: Complete directed acyclic graphs are Markov equivalent and represented by an essential graph with undirected edges only.

Directed acyclic graphs have often been treated in the literature on graph theory. The hybrid essential graphs have a rather special structure that has only been studied in the literature on graphical models. For the enumeration of some classes of conditional independence models there exist formulas and/or algorithms [8, 5, 6, 7, 9, 10]. Some of our own results were reported at previous workshops.

A number of related questions arise in the field of constraint programming. The search space of matrix optimisation problems may be unnecessarily large because permutations of rows and columns have the same solutions and thus build an equivalence class. Before applying constraint programming the symmetries are broken by lexicographically ordering the matrix. Several results from the literature on constraint programming are relevant to the present problem [4] .

To my knowledge for the enumeration of essential graphs no satisfactory solutions exists. The goal of the present contribution is not just to develop an enumeration algorithm but to enumerate essential graphs according to important structural properties.

We consider only graphs with one component. Results for graphs with two or more components may be obtained by combining several one-component analyses. The paper is mainly concerned with the enumeration of unlabeled

structures. We will see below that the enumeration of the labelings of a given essential graph is easy. Finding the unlabeled structures is the main problem.

2 Generalized terminal nodes

In a directed graph a terminal node is a vertex without children. We generalize the concept to essential graphs. In an essential graph a vertex is a generalized terminal if its undirected links can be replaced by directed ones so that there exists a Markov equivalent graph in which the vertex is an ordinary terminal.

Definition 1 (Generalized terminal) *Let $G = (V, E)$ be an essential graph with vertex set V and edges E, and let $\mathcal{M}(G) = \{DAG_1, \ldots, DAG_t\}$ be the set of its Markov equivalent DAGs. $X_i \in V$ is a generalized terminal vertex if there exists a $DAG \in \mathcal{M}(G)$ in which X_i is an (ordinary) terminal.*

In Figure 1 the verteces B and C are ordinary terminals in (a) and generalized terminals in (d). The graph in (a) is Markov equivalent to the graph in (d). Therefore B and C are generalized terminals in (d). In panel (g) of Figure 2 the verteces A, B, and C are generalized terminals as for each of them there exist a Markov equivalent directed graph in which it is an ordinary terminal. In a complete graph all verteces are generalized terminals. Likewise, all the verteces in an edgeless graph are generalized terminals.

2.1 Layering

We use the generalized terminal verteces to represent an essential graph as a *layered graph*. The last layer consists of the terminals of the given graph. We remove the terminals and determine the next layer by the terminals of the remaining subgraph. We continue iteratively till the remaining graph is empty. The layers introduce a partial order on the verteces.

The number of layers and the according number of verteces within each of the layers is called the *layering* of the graph. It is written $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$, where m denotes the number of layers and λ_i the number of verteces in layer i. Each layering is a subdivision the sequence of n elements (see Table 1).

An easy and efficient way to generate the subdivisions is to work with a binary encoding of the elements in each subset. The full set of elements corresponds to the bit representation of $2^n - 1$, where n is the number of verteces. Throughout we number the verteces from 0 instead of 1.

The number of subdivisions with m subsets is obtained by the binomial coefficients $\binom{n}{m}$. The total number of different subdivisions is

$$\sum_{m=0}^{n} \binom{n}{m} = 2^{n-1} \tag{1}$$

A layering is not admissible if each of its first two layers contains one element only. In this case the first element would be a generalized terminal node, and this contradicts the definition of layers. Layers of the type $\lambda = (1, 1, \ldots, \lambda_m)$ are non-admissible. The number of layerings that are excluded by this constraint is 2^{n-1-2}. We thus obtain the following result.

	Binary encoding	Subdivision	Number of layers
1	31	$\{0,1,2,3,4\}$	1
2	15, 16	$\{0,1,2,3\},\{4\}$	2
3	7, 24	$\{0,1,2\},\{3,4\}$	2
4	7, 8, 16	$\{0,1,2\},\{3\},\{4\}$	3
5	3, 28	$\{0,1\},\{2,3,4\}$	2
6	3, 12, 16	$\{0,1\},\{2,3\},\{4\}$	3
7	3, 4, 24	$\{0,1\},\{2\},\{3,4\}$	3
8	3, 4, 8, 16	$\{0,1\},\{2\},\{3\},\{4\}$	4
9	1, 30	$\{0\},\{1,2,3,4\}$	2
10	1, 14, 16	$\{0\},\{1,2,3\},\{4\}$	3
11	1, 6, 24	$\{0\},\{1,2\},\{3,4\}$	3
12	1, 6, 8, 16	$\{0\},\{1,2\},\{3\},\{4\}$	4
13	1, 2, 28	$\{0\},\{1\},\{2,3,4\}$	3
14	1, 2, 12, 16	$\{0\},\{1\},\{2,3\},\{4\}$	4
15	1, 2, 4, 24	$\{0\},\{1\},\{2\},\{3,4\}$	4
16	1, 2, 4, 8, 16	$\{0\},\{1\},\{2\},\{3\},\{4\}$	5

Table 1: The 12 admissible layering for $n = 5$ (row 1 to 12) and the four non-admissible structures (row 13 to 16)

Theorem 1 (Number of layerings) *The number of layerings of an essential graph with n nodes is*

$$\ell = 3 \cdot 2^{n-3} \,. \tag{2}$$

3 Symmetry

Using the layering of an essential graph we introduce the following definition of symmetric verteces.

Definition 2 (Symmetry) *Two verteces of an essential graph are symmetric*

1. *if they both are in the same layer and have the same number of neighbors in this layer, and*
2. *if all their neighbors are pairwise symmetric.*

Sets of mutually symmetric verteces build symmetry classes, $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_q)$, where σ_1, σ_2, $\ldots$, σ_q denote the cardinalities of the classes. The classes of symmetric verteces are used to determine the number of possible labelings of a given essential graph.

Theorem 2 (Number of labelings of an essential graph) *The number of labelings of an essential graph with q symmetry classes with cardinalities $\sigma_1, \ldots, \sigma_q$ is*

$$\mathcal{L} = \frac{n!}{\sigma_1!\ \sigma_2!\ \cdots \sigma_q!} \,. \tag{3}$$

If there are no symmetries, then there are $n!$ different labelings. If σ_1 of them are indistinguishable, then there are $\sigma_1!$ permutations that count only as one pattern so that there are only $\frac{n!}{\sigma_1!}$ distinguishable labelings etc. for the remaining symmetry classes.

	0	1	2	3	4	5	6	7
0								
1	28							
2	27	14						
3	9	26	25					
4	8	24	23	13				
5	7	22	21	12	11			
6	2	6	5	20	19	18		
7	1	4	3	17	16	15	10	

Table 2: Ranks of the cells of an adjacency matrix with layering $\lambda = (1, 2, 3, 2)$

4 Representatives

A set of Markov equivalent Bayesian networks may be characterised by selecting just one representative. It represents its equivalance class. We define the representative by the order of the verteces. The order between the layers is fixed. The order within the layers needs is free. The adjacency matrix (with its 0/1 entries) may be conceived as one big binary number. We select the representative by the maximum of this binary number. All $n!$ permutations would be needed to find the maximum order by brute force methods. The layers reduce the computations to within-layer permutations. But we will improve the procedure much further by exploiting the layering and the symmetries. The advantage of the use of such structural properties is that they will enable us to enumerate interesting subclasses of models.

We assign a rank r_{ij} to each cell in the adjacency matrix, $1 \leq r_{ij} \leq n(n-1)/2$, and take the sum of $2^{r_{ij}}$ values as our criterion,

$$C = \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} 2^{r_{ij}}. \tag{4}$$

The integer C encodes the binary number that corresponds to the adjacency matrix. Table 2 shows a matrix for the layering $\lambda = (1, 2, 3, 2)$.

Which cell of the adjacency matrix obtains which rank is our choice. We assign the highest ranks to the cells that correspond to the edges between two neighboring layers. These edges determine the layering. If all other edges are removed from the graph the layering does not change. Moreover, there is a minimum assignment of edges that defines the layering. All edges in this substructure connect layers with distance $d = 1$. Edges at the beginning of the layers obtain the highest ranks.

The second series of ranks is assigned to within-layer edges. They build a band along the diagonal of the adjacency matrix and have the layer distance $d = 0$. Finally, we assign ranks to cells with distances $d = 2, 3$ etc. up to the largest distance between the layers. The cell in the lower left corner has always rank 1. The powers of 2 criterion can always be expressed by a specific lexicographical ordering.

5 Backbone

To investigate the unlabeled structures we introduce the concept of a *backbone*.

Definition 3 (Backbone) *Let G be an essential graph with the layer structure λ. A subgraph of G that is induced by the set of those edges that connect verteces in two adjacent layers, is called a backbone of G.*

Figure 3 shows the twenty backbones for $n = 5$. All edges in the backbone are directed edges, with one exception: edges between the first and the second layer may be directed or undirected. Panel (a) of Figure 4 shows an example. For $n \geq 8$ the backbone may consist of several components.

5.0.1 Minimal backbone

Each layering induces a *minimal backbone*. The minimal backbone is used in an algorithm that generates all essential graphs with a given layer structure. In the adjacency matrix the edges connecting two layers build a rectangular submatrix, $L_u \times L_v$, where L_v is the parent and L_u, $u > v$, the child layer. We call such a submatrix a tableau and denote it by T_{uv}. For $d > 0$ each tableau corresponds to a bipartite graph.

Two successive layers impose a minimum pattern of 1s in the tableau or edges in the graph, respectively. Each vertex in a row, v_i, must be a generalised terminal. It must get at least one link from the preceding layer. Likewise, each vertex in a column must have at least one link to the next layer. Otherwise it would be unconnected and be a generalised terminal belonging to the next layer. We have seen that the first two layers are special. We have the following properties.

Properties 1 (Minimal backbone) *A minimal backbone is characterised by the following lexicographical order:*

1. *In the case of non-initial layers:*
   (a) *If T_{uv}, $u = 3, \ldots, m, v = u - 1$, is quadratic ($\lambda_u = \lambda_v$), then the minimum assignment has 1s in the main diagonal and 0s in all other entries.*
   (b) *If T_{uv} is non-quadratic ($\lambda_u \neq \lambda_v$), then the minimum assignment has 1s in the main diagonal and, in addition, $\lambda_v - \lambda_u$ ($\lambda_u - \lambda_v$) 1s in the first row (in the first column).*
2. *In the case of initial layers: If T_{uv}, $u = 2, \ldots, m, v = 1$, then for each 1 there is a second 1 in the same row or the same column and the tableau is the lexicographically weakest among all tableaux with this property.*

Here are three examples of minimum non-initial backbone tableaux:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

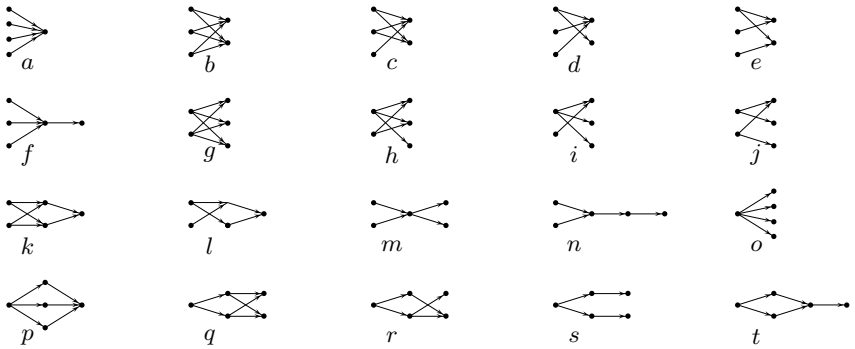


Figure 3: Backbones for $n = 5$; $\{b, c, d, e\}, \{g, h, i, j\}, \{k, l\}$, and $\{q, r, s\}$ have the same layering; e, j, l, and s are their minimal backbones

The rows and columns are lexicographically ordered (top $\geq$ down, left $\geq$ right). Because of the vee/wedge condition the minimum assignment for $T_{2,1}$ is different. Each 1 in the matrix must have a second 1 in the same row or in the same column and the assignment must be a lexicographical minimum. Here three examples of minimum initial backbone tableaux:

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Again, the rows and columns are lexicographically ordered. In addition in the first layer each vertex must be the origin of two or more diverging edges, or it must be the origin of an edge that converges with other edges in the second layer. Each vertex must be part of a vee or a wedge. Otherwise it would be a generalised terminal and not bee in first layer.

5.1 Backbone criteria

All possible backbones are obtained by replacing the 0s in the tableaux by 1s. The minimal backbones are a good starting point. It would not be correct, though, to fill in 1s for 0s in all possible combinations. We have to test each possible tableau by the following *backbone criteria*:

Properties 2 (Backbone) *An adjacency matrix represents a backbone structure iff all its tableaux with $d = 1$ have the following properties:*

1. *All column and row sums are > 0.*
2. *The column and row sums are weakly decreasing top $\geq$ down and left $\geq$ right.*
3. *If two or more rows have equal sums, then they are weakly lexicographically ordered top down.*
4. *In the first layer each node is a member of a vee or a wedge, i.e., $\lambda = (1, 1, \ldots)$ layerings are non-admissible.*

These criteria are used to generate all possible backbones. Table 3 shows counts up to $n = 12$ verteces. We emphasise that the counts are much smaller than the counts for all possible models. For model learning it seems possible to search for the best fitting backbone(s) or minimal backbone(s). A hierarchical search method mays often be reasonable because "direct causes" usually are more interesting than "long distance causes" or undirected within-layer connections.

n	Backbones	n	Backbones
4	6	9	4.694
5	20	10	23.577
6	69	11	133.626
7	256	12	868.034
8	1.042		

Table 3: Number of backbone structures for essential graphs with n verteces.

5.2 Within-layer edges

We next consider the edges connecting verteces within a layer. In the adjacency matrix the within-layer edges are located in a band of triangles along the main diagonal. These edges are also constrained by the layer structure.

Theorem 3 (Within-layer edges) *Within each layer all edges belong to completely connected components and are undirected.*

If there would be a vee-structure of the type $X \to Y \leftarrow Z$, then the vertex Y would be a generalized terminal and belong to the next layer. Correspondingly, if there would be a wedge-structure of the type $X \leftarrow Y \to Z$, then X and Z would be generalized terminals and belong to the next layer. Only if X, Y and Z are completely interconnected, are they generalized terminals and belong to the same layer. This holds of course also for more than three verteces. Thus, the within-layer verteces must have no links at all (in this case each isolated vertex is an extreme form of a complete component) or they must belong to a completety connected component. As a consequence, all edges within one layer are undirected.

The components build a partition of the verteces in a layer. The number of partitions of k is denoted by $p(k)$. It may be calculated by an algorithm given, e.g., in [2]. In a compute program it is more efficient to store the numbers (see Table 4) in an array. We combine the number of layerings and the number of partitions within each layer to obtain the number of within-layer structures N_w.

Theorem 4 (Number of within-layer structures) *The number of within-layer structures in all layerings is*

$$N_w = \sum_{k=0}^{n-1} \binom{n}{k} p(k)\,. \tag{5}$$

In addition, we combine the enumeration of the backbones with the formula of within-layer structures to obtain the counts for $n = 5$ shown in Table 5. These

k	$p(k)$	k	$p(k)$	k	$p(k)$	k	$p(k)$
1	1	6	11	11	56	16	231
2	2	7	15	12	77	17	297
3	3	8	22	13	101	18	385
4	5	9	30	14	135	19	490
5	7	10	42	15	176	20	627

Table 4: Number of partitions up to $n = 20$. Taken from Andrews [2, Table 14.1]

Partition	Nbr of Backbones	Nbr of within-layer patterns	Total	
1	$\{0,1,2,3,4\}$	-	1	1
2	$\{0,1,2,3\},\{4\}$	1	5×1	4
3	$\{0,1,2\},\{3,4\}$	4	3×2	16
4	$\{0,1,2\},\{3\},\{4\}$	1	$3 \times 1 \times 1$	8
5	$\{0,1\},\{2,3,4\}$	4	2×3	14
6	$\{0,1\},\{2,3\},\{4\}$	2	$2 \times 2 \times 1$	18
7	$\{0,1\},\{2\},\{3,4\}$	1	$2 \times 1 \times 2$	14
8	$\{0,1\},\{2\},\{3\},\{4\}$	1	$2 \times 1 \times 1 \times 1$	12
9	$\{0\},\{1,2,3,4\}$	1	1×5	4
10	$\{0\},\{1,2,3\},\{4\}$	1	$1 \times 3 \times 1$	6
11	$\{0\},\{1,2\},\{3,4\}$	3	$1 \times 2 \times 2$	12
12	$\{0\},\{1,2\},\{3\},\{4\}$	2	$1 \times 2 \times 1 \times 1$	10
			119	

Table 5: Number of one-component structures, $n = 5$, with parents or children in the previous or next layer ($d = 1$) and with links within the same layer ($d = 0$). Row 1 corresponds to the complete graph.

counts contain all models with $d = 1$ (backbones) and $d = 0$ (within-layer components).

5.3 Long distance edges

We procede to edges which connect verteces in non-adjacent layers, i.e., edges with layer distance $d > 1$. All these edges are directed. Lexicographical order is used to break the symmetries (as defined in definition 2) in long distance edges.

If a graph has m layers, we considere tableaux with parents in layer L_k and their children in L_{k+d}, $d = 2, \ldots, m-2$. In the adjacency matrix these tableaux build a band of rectangles parallel to the main diagonal with distance d.

Maximization seems to require simultaneous lex ordering of rows and columns. This is not true. We consider tableaux with *symmetric* parents and *symmetric* children only.

We test the following *long-distance criteria*:

Properties 3 (Long-distance) *Long-distance symmetries are broken by lex ordering.*

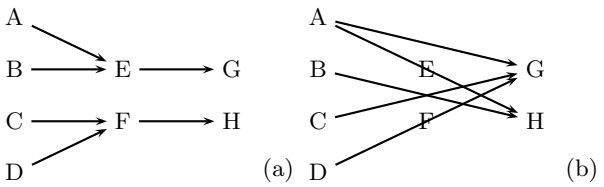


Figure 4: (a) Backbone with symmetry classes $\{A, B, C, D\}$, $\{E, F\}$, and $\{G, H\}$. (b) Example of admissible arcs from layer 1 to layer 3.

1. *Find the symmetries in the adjacency matrix above the current row i (but including the backbone, the within-layer edges, and already inserted long-distance edges),*

2. *If there are two (or more) symmetric verteces in the input layer of vertex v_i (row i), say v_j and v_k (two symmetric columns), then $a_{ij} \geq a_{ik}$.*

Finding the symmetries for each row is computationally expensive.

There is a special class of models for which these criteria hold. All tableaux which have the form of a Young tableau fulfill these requirements. Here the first child gets inputs from the first p parents, the second child gets input from the first q parents, where $q \leq p$, and so on. The first elements of a symmetry class may always be set to 1.

We have written a C program that combines the steps described in the previous sections. The program enumerates the essential graphs for a given number of verteces.

6 Discussion

The structure of graphical models is extracted from (case I) experts or from (case II) statistical data. The two alternatives reflect two extreme attitudes toward the introduction of *prior information*. In the first case an expert assesses the qualitative structure of the model. The data are only used to estimate the numerical probabilities. This imposes a heavy load upon the prior knowledge of the expert. In the second alternative "no" prior knowledge is invoked. This is done by giving each possible structure the same plausibility.

Using a uniform distribution over the set of Bayesian networks is problematic because Markov equivalent models lead to multiple counts. Bayesian networks with many equivalent models obtain higher probabilities so that the distribution is in fact not uniform. Multiple counts lead to incoherence.

Case I and case II are extreme endpoints on a continuum. We usually have *some* knowledge about the structure of the model. We often think that the variables we investigate belong to one component. We may start our analysis by testing this assumption. We next may investigate the layering of the model. The backbones are important as we often have rather strong intuitions about direct influences. We note that edges linking the first two layers a special. They must satisfy the vee/wedge condition. In the wedge condition they contain undirected

edges. In a stepwise procedure it is reasonable to start at the "beginning" of the backbone and add layers one by one. Darwiche [3, p. 456ff.] discusses similar steps fitting tree structures.

Adding undirected edges is the next step. Finally, long distance influences may be the last step in the analysis. Such a stepwise method leads to a hierachical analysis. Both, the prior distribution and the learning of model structures is guided by top-down procedures, starting with direct "causes" and proceding to more indirect long distance dependencies.

References

[1] Andersson, S. A., Madigan, D., and Perlman, M. D. (1998). A characterization of Markov equivalance classes for acyclic digraphs. Annals of Statistics, **27**, 502-541.

[2] Andrews, G. E. (1984). *The Theory of Partitions*. Cambridge University Press. Cambridge, UK.

[3] Darwiche, A. (2009). Modeling and Reasoning with Bayesian Networks. Cambridge University Press, Cambridge, New York.

[4] Flener, P., Frisch, A., Hnich, B., Kiziltan, Z., Miguel, I., Pearson, J., and Walsh, T. (2002). Breaking row and column symmetries in matrix models. In P. Van Hentenryck (Ed.), *Constraint Programming*, LNCS 2470, 462-477.

[5] Gillispie, S. B. and Perlman, M. D. (2001). Enumerating Markov equivalence classes of acyclic digraph models. In Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference , Morgan Kaufmann, San Francisco, 171-177.

[6] Gillispie, S. B. and Perlman, M. D. (2002). The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence*, **141**, 137-155.

[7] Gillispie, S. B. (2003). Formulas for counting acyclic digraph Markov equivalence classes. Technical report. Department of Radiology, University of Washington.

[8] Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. *Lecture notes Mathematics*, **622**, 28-43.

[9] Steinsky, B. (2003) Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete Mathematics*, **270**, 267-278.

[10] Steinsky, B. (2003) Efficient coding of labeled directed acyclic graphs. *Soft Computing*, **7**, 350-356.

Completions of Fragments of Lattice-Valued Possibilistic Distributions According to the Principle of Maximum Entropy Value

Ivan Kramosil
Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Prague 8,
Czech Republic
fax: (+420) 268 585 789, e-mail: kramosil@cs.cas.cz

Abstract

Investigated are Boolean-valued possibilistic distributions taking their values in the power-set of all sets of positive integers. However, some of these possibility degrees may be known only fragmentally in the sense that for the characteristic sequence (identifier) of the set-value in question not all members of this sequence are known. A simple possibilistic entropy function is defined and completions of fragments of possibility degrees with respect to the classical (optimistic or global, in a sense) principle of maximal entropy as well as with respect to some weakened (local or pessimistic, in a sense) versions of this principle are introduced and analyzed.

1 Introduction and Motivation

The notions of possibilistic distribution and possibilistic measure were conceived by L. A. Zadeh in [13] rather as an alternative description of the notions from the theory of fuzzy sets in terms syntactically more similar to those used in measure theory in general and in probability theory in particular. Consequently, also the shift from real-valued fuzzy sets to their non-numerical and, in particular, lattice-valued modifications, has been taken as an inspiration for the investigation of lattice-valued possibilistic distributions and measures, i.e., mappings ascribing to elementary events and their sets values from a complete lattice, so quantifying the degree of possibility ascribed to this element or set. The reader should consult [5], the excellent monograph [2], or some other relevant source when also the reasons leading to the choice of complete lattice as the structure over uncertainty (possibility) degrees are analyzed in more detail [3].

In this paper we will investigate, which of the properties possessed by lattice-valued possibilistic distributions and measures remain (or do not remain) to be valid, if the conditions imposed on the structure of values of possibility degrees are weakened (e.g., in [9] we analyzed the case when this structure defines

just a lattice, i.e., not necessary a complete one). Namely, in what follows, we analyze Boolean-valued possibilistic distributions and measures such that possibility degrees are quantified by sets of positive integers. Hence, possibilistic distributions will be defined by mappings π which take a nonempty space Ω into the power-set $\mathcal{P}(\mathcal{N})$ over the set $\mathcal{N} = \{1, 2, \ldots\}$ of positive integers and such that $\bigcup_{\omega \in \Omega} \pi(\omega) = \mathcal{N} = \{1, 2, \ldots\}$ holds. However, our situation will be put into serious troubles by the assumption that the values $\pi(\omega)$ are not, in general, completely known or identifiable. Replacing each $A \subset \mathcal{N}$ by its identifier (characteristic function) $\chi(A) : \mathcal{N} \to \{0, 1\}$, we have to admit that for some $\omega \in \Omega$ and some $i \in \mathcal{N}$ the values $\chi(\pi(\omega))$ (i) are not known and must be replaced by some abstract auxiliary symbol λ. Hence, instead of Boolean-valued ($\langle \mathcal{P}(\mathcal{N}), \subseteq \rangle$-valued, more precisely) mapping $\pi : \Omega \to \{0, 1\}^\infty$ we have at hand just a fragment of π defined by a mapping $\pi^* : \Omega \to \{0, 1, \lambda\}^\infty$.

Inspired by the classical Shannon entropy function (cf. [6], e.g.), and replacing the integration with respect to probability measure by Sugeno integral, we arrive at a very simple possibilistic entropy function and we prove easily that the completion of sequences from $\{0, 1, \lambda\}^\infty$ by replacing all the occurrences of λ by 1 meets the principle of maximum possibilistic entropy function imposed on the mappings from $\{0, 1, \lambda\}^\infty$ to $\{0, 1\}^\infty$. Applying to these mappings some "more pessimistic" or "safety first" demands, we propose and analyze mappings $\{0, 1, \lambda\}^\infty \to \{0, 1\}^\infty$ such that occurrences of λ are replaced by 1 without touching the occurrences of 0 and 1 in the original $\{0, 1, \lambda\}^\infty$-sequence only if the set of all occurrences of λ meets some restrictions, e.g., the set of such occurrences in "small enough", they are localized "close to each other", the set of all occurrences of λ can be covered by a "small" or "easy to define" subset of $\mathcal{N} = \{1, 2, \ldots\}$, etc., cf. below for more detail.

The paper is written on an almost self-explanatory level, just some rather elementary preliminaries on partial orderings, lattices, Boolean algebras and some other structures are assumed; the reader may consult [1, 4] or [11] (or some more recent textbook or monograph) for these sakes.

2 Lattice-Valued Entropy Functions and the Principle of Maximum Uncertainty

Let us propose, given a mapping $\pi^T : \Omega \to \{0, 1, \lambda\}^\infty$, a method how to embed this mapping into $\{0, 1\}^\infty$, in other terms, how to replace all the occurrences of λ in $\{\pi^T(\omega) : \omega \in \Omega\}$ by 0 or 1 in a way which could be taken as optimal according to a reasonable criterion. As such a criterion we apply a lattice-valued modification of the principle of maximum entropy (in the case of probability measures) or the principle of maximum uncertainty (in the case of other numerical quantifications of uncertainty). Cf. [6] as an excellent survey of works dealing with various models of uncertainty quantification and processing.

Let $\leq_0$ be the partial ordering on $\{0, 1\}^\infty$ defined in such a way that for $\langle x_1, x_2, \ldots \rangle, \langle y_1, y_2, \ldots \rangle \in \{0, 1\}^\infty$, $\langle x_1, x_2, \ldots \rangle \leq_0 \langle y_1, y_2, \ldots \rangle$ holds iff $x_i \leq y_i$ holds for each $i \in \mathcal{N}$ and $0 < 1$ holds on $\{0, 1\}$. The binary relation $\leq_0$ obviously meets the conditions imposed on partial ordering and $\langle \{0, 1\}^\infty, \leq_0 \rangle$ defines a complete lattice on $\{0, 1\}^\infty$. The explicit formulas for supremum ($\bigvee^0$) and infimum ($\bigwedge^0$) operations induced by $\leq_0$ on $\{0, 1\}^\infty$ are worth being

introduced explicitly. Writing $\boldsymbol{x} = \langle (\boldsymbol{x})_1, (\boldsymbol{x})_2, \ldots \rangle$ for $\boldsymbol{x} \in \{0,1\}^\infty$, we obtain, for each $\emptyset \neq B \subset \{0,1\}^\infty$, that

$$\vee^0 B = \bigvee_{\boldsymbol{x} \in B}^{0} \boldsymbol{x} = \left\langle \bigvee_{\boldsymbol{x} \in B} (\boldsymbol{x})_i \right\rangle_{i=1}^{\infty}, \tag{2.1}$$

$$\wedge^0 B = \bigwedge_{\boldsymbol{x} \in B}^{0} \boldsymbol{x} = \left\langle \bigwedge_{\boldsymbol{x} \in B} (\boldsymbol{x})_i \right\rangle_{i=1}^{\infty}, \tag{2.2}$$

where $\vee$ and $\wedge$ denotes the supremum and infimum operation on $\{0,1\}$, induced by the relation $0 < 1$.

A similar formalized structure over the space $\{0,1,\lambda\}^\infty$ of infinite ternary sequences may be conceived as follows. Let $\leq$ be the linear ordering on $\{0,1,\lambda\}$, completely defined by the relation $0 < \lambda < 1$. Partial ordering $\leq_T$ on $\{0,1,\lambda\}^\infty$ is defined so that, for each $\boldsymbol{x} = \langle x_1, x_2, \ldots \rangle, \boldsymbol{y} = \langle y_1, y_2, \ldots \rangle \in \{0,1,\lambda\}^\infty, \boldsymbol{x} \leq_T \boldsymbol{y}$ holds iff $x_i \leq y_i$ is valid for each $i \in \mathcal{N} = \{1,2,\ldots\}$. The pair $\langle \{0,1,\lambda\}^\infty, \leq_T \rangle$ obviously defines a complete lattice over $\{0,1\langle\}^\infty$, hence, supremum $\bigvee^T$ and infimum $\bigwedge^T$ operations are defined for each $\emptyset \subset \{0,1\lambda\}^\infty$ and their explicit definitions are identical with (2.1) and (2.2), just keeping in mind that $\bigvee_{\boldsymbol{x} \in B} (\boldsymbol{x})_i$ and $\bigwedge_{\boldsymbol{x} \in B} (\boldsymbol{x})_i$ are defined over the set $\{0,1,\lambda\}$, not only over $\{0,1\}$ as it is the case in (2.1) and (2.2).

When seeking for a lattice-valued modification of an entropy or uncertainty function applicable to possibilistic distributions $\pi : \Omega \to \{0,1\}^\infty$ we take inspiration from the classical Shannon entropy function H. Given a finite or countable space $\Omega = \{\omega_1, \omega_2, \ldots\}$ and a probability distribution p on Ω, i.e., $p : \Omega \to [0,1]$ is such that $\Sigma_{i=1}^{\infty} p(\omega_i) = 1$, then Shannon entropy H of p is defined by

$$H(p) = -\sum_{i=1}^{\infty} p_i \log_2(p_i) = \sum_{i=1}^{\infty} p_i \log_2(1/p_i) = \sum_{\omega \in \Omega} p(\omega) \log_2(1/p(\omega)). \tag{2.3}$$

Hence, $H(p)$ is defined as the expected value of the decreasing function $\log_2(1/p(\omega))$ of $p(\omega)$. Replacing this function by another nonincreasing function of $p(\omega)$, namely by the function $1-p(\omega)$, we arrive at the function $\sum_{\omega \in \Omega} p(\omega) P(\Omega - \{\omega\})$, where P is the probability measure on $\mathcal{P}(\Omega)$ induced by p (cf. [10] and [12] for more detail). In order to shift our model from probability to possibility theory let us replace $P(\Omega - \{\omega\})$ by $\Pi(\Omega - \{\omega\}) = \bigvee_{\omega_0 \in \Omega, \omega_0 \neq \omega} \pi(\omega_0)$, $\sum_{\omega \in \Omega}$ by $\bigvee_{\omega \in \Omega}^{0}$, and product by infimum $\wedge_0$ in the complete lattice $\langle \{0,1\}^\infty, \leq_0 \rangle$, so that we arrive at the Sugeno integral (cf. [2], [8]) ascribing to $\langle \{0,1\}^\infty, \leq_0 \rangle$-valued possibilistic distribution π on Ω the entropy or uncertainty value

$$I(\pi) = \bigvee_{\omega \in \Omega}^{0} [\pi(\omega) \wedge_0 \Pi(\Omega - \{\omega\})]. \tag{2.4}$$

This quantification of uncertainty is not too fine or flexible when $I(\pi) = 1^\infty$ is the case. Indeed, if $\pi(\omega_1) = \pi(\omega_2) = 1^\infty$ for $\omega_1, \omega_2 \in \Omega, \omega_1 \neq \omega_2$, then for each $\omega \in \Omega$ either ω_1 or ω_2 is in $\Omega - \{\omega\}$, so that $\Pi(\Omega - \{\omega\}) = 1^\infty$ and

$$I(\pi) = \bigvee_{\omega \in \Omega}^{0} [\pi(\omega) \wedge_0 \Pi(\Omega - \{\omega\}) = \bigvee_{\omega \in \Omega}^{0} \pi(\omega) = 1^\infty \tag{2.5}$$

holds (a refinement of this lattice-valued entropy function can be found in [7]). Nevertheless, let us apply $I(\pi)$ as our first attempt to apply the maximum entropy principle when completing the missing items in the $\pi^T : \Omega \to \{0,1,\lambda\}^\infty$ mappings.

Lemma 2.1 *Let $\pi_1, \pi_2 : \Omega \to \{0,1\}^\infty$ be two mappings, let $\pi_1(\omega) \leq_0 \pi_2(\omega)$ hold for each $\omega \in \Omega$. If π_1 defines a $\langle\{0,1\}^\infty, \leq_0\rangle$-valued possibilistic distribution on Ω, the same is the case for π_2 and the relation $I(\pi_1) \leq_0 I(\pi_2)$ holds.*

Proof: If $\bigvee_{\omega\in\Omega}^\infty \pi_1(\omega) = 1^\infty$ and $\pi_1(\omega) \leq_0 \pi_2(\omega)$ for each $\omega \in \Omega$ holds, then $\bigvee_{\omega\in\Omega}^0 \pi_2(\omega) = 1^\infty$ follows, so that π_2 defines a $\langle\{0,1\}^\infty, \leq_0\rangle$-valued possibilistic distribution on Ω. For each $A \subset \Omega, \Pi_1(A) = \bigvee_{\omega\in A}^0 \pi_1(\omega) \leq \bigvee_{\omega\in A}^0 \pi_2(\omega) = \Pi_2(A)$ holds, in particular, $\Pi_1(\Omega - \{\omega\}) \leq_0 \Pi_2(\Omega - \{\omega\})$ holds for each $\omega \in \Omega$, so that the inequality

$$I(\pi_1) = \bigvee_{\omega\in\Omega}^0 [\pi_1(\omega) \wedge_0 \Pi_1(\Omega - \{\omega\})] \leq_0 \bigvee_{\omega\in\Omega}^0 [\pi_2(\omega) \wedge_0 \Pi_2(\Omega - \{\omega\})] = I(\pi_2) \quad (2.6)$$

follows and the assertion is proved. □

Let $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$ be an infinite ternary sequence possibly containing some occurrence(s) of λ, let $\boldsymbol{x}^+ \in \{0,1\}$ be defined by replacing all λ's in $\boldsymbol{x}$ by 1. Obviously, for each $\boldsymbol{y} \in \{0,1\}^\infty$ such that $(\boldsymbol{y})_i = (x)_i$ supposing that $(\boldsymbol{x})_i \in \{0,1\}$ holds, the inequality $\boldsymbol{y} \leq_0 \boldsymbol{x}^+$ follows. Hence, given a $\langle\{0,1,\lambda\}^\infty, \leq_T\rangle$-valued possibilistic distribution π on Ω, and setting $\pi^0(\omega) = (\pi(\omega))^+$ for each $\omega \in \Omega$, we obtain easily that $I(\pi^*) \leq_0 I(\pi^+)$ holds for each $\langle\{0,1\}^\infty, \leq_0\rangle$-valued possibilistic distribution π^* on Ω such that $(\pi^*(\omega))_i = (\pi(\omega))_i$ for each $\omega \in \Omega$ and each $i \in \mathcal{N} = \{1,2,\ldots\}$, for which $\pi(\omega))_i \neq \lambda$.

3 Local and Global Principle of Maximum Uncertainty

Lemma 2.1 offers the most simple way how to embed sequences from $\{0,1,\lambda\}^\infty$ into $\{0,1\}^\infty$ according to the maximum entropy or uncertainty principle – to replace each sequence $\boldsymbol{x} \in \{0,1,\lambda\}$ by the binary sequence $\boldsymbol{x}^+$. Hence, all the values 0 and 1 in $\boldsymbol{x}$ are taken as sure and reliable. Let us introduce, in this section a generalized version of this approach.

Let $\mathcal{S}$ be a system of subsets of $\mathcal{N} = \{1,2,\ldots\}$, let $\boldsymbol{x} = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\rangle \in \{0,1,\lambda\}^\infty$, let $F(\boldsymbol{x}) = \{i \in \mathcal{N} : \boldsymbol{x}_i = \lambda\}$. Then $\varphi(\mathcal{S})(\boldsymbol{x})$ is the binary sequence from $\{0,1\}^\infty$ defined in this way: if there exists $S \in \mathcal{S}$ such that $F(\boldsymbol{x}) \subset S$ holds, then $\varphi(\mathcal{S})(\boldsymbol{x}) = \boldsymbol{x}^+$, otherwise $\varphi(\mathcal{S})(\boldsymbol{x}) = 1^\infty$. Let us emphasize that the set $F(\boldsymbol{x})$ must be covered by *one* set $S \in \mathcal{S}$ in order to apply the transformation $\varphi(\mathcal{S})(\boldsymbol{x}) = x^+$.

Let us prove that the sequence $\varphi(\mathcal{S})(\boldsymbol{x})$ is defined uniquely, even if there may be two or more sets in $\mathcal{S}$ covering the set $F(\boldsymbol{x})$. Indeed, if $\mathcal{S}_0 = \{S_0 \in \mathcal{S} : F(\boldsymbol{x}) \subset S_0\}$, then $F(\boldsymbol{x}) \subset \bigcap \mathcal{S}_0$ holds. Hence, $\boldsymbol{x}_i = 0$ or 1 for every $S_0 \in \mathcal{S}_0$ and every $i \in S_0 - \bigcap \mathcal{S}_0$, so that $(\varphi(\mathcal{S})(\boldsymbol{x}))_i = \boldsymbol{x}_i$, if $i \in S - \bigcap \mathcal{S}_0 (\varphi(S)(\boldsymbol{x}))_i = 1$, if $i \in \bigcap \mathcal{S}_0$ holds, no matter which $S \in \mathcal{S}$, $F(\boldsymbol{x}) \subset S$ is applied. Consequently, if

there are sets $S_1, S_2 \in \mathcal{S}$, $S_1 \subset S_2, S_1 \neq S_2$, then $\varphi(\mathcal{S} - \{S_1\})(\boldsymbol{x}) = \varphi(\mathcal{S})(\boldsymbol{x})$ for each $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$.

An interpretation behind the mapping $\varphi(\mathcal{S}) : \{0,1,\lambda\}^\infty \to \{0,1\}^\infty$ may read as follows. Investigating a sequence $\pi(\omega) = \langle(\pi(\omega))_i\rangle_{i=1}^{\infty}\rangle$ with some missing components, we can separate two cases. Either we know, because of some preliminary knowledge concerning the source producing the data in question, that all indices $i \in \mathcal{N}$ for which $(\pi(\omega))_i$ is not known are of such kind that they do not influence correctness of the values $(\pi(\omega))_j$ for indices out of $F(\boldsymbol{x})$, i.e., for indices for which $(\pi(\omega))_j$ is 0 or 1. Consequently, when approximating the missing values under the principle of maximum uncertainty, we may limit ourselves just to the missing values keeping the known (i.e., $0-1$ values) intouched. It is just the case when the set of indices with missing values is covered by some set from $\mathcal{S}$. On the other side, for some indices not coverable by a set from $\mathcal{S}$ the failure to identity the value $(\pi(\omega))_i$ may announce a possible error when identifying the values $(\pi(\omega))_j$ for some other $j \in \mathcal{N}$, even when we obtained, somehow, that $(\pi(\omega))_j$ is 0 or 1. Hence, all the processed sequence $\pi(\omega) \in \{0,1,\lambda\}^\infty$ is doubtful and does not contain any sure information about the original binary sequence. Consequently, all the sequences from $\{0,1\}^\infty$ must be re-considered when approximating the sequence from $\{0,1,\lambda\}^\infty$, hence, it is only the sequence $1^\infty = \langle 1,1,\ldots\rangle \in \{0,1\}^\infty$ which maximizes the entropy $I(\pi)$ for π ranging over $\{0,1\}^\infty$. So, the application of the mapping $\varphi(\mathcal{S})$ when projecting $\{0,1,\lambda\}^\infty$ into $\{0,1\}^\infty$ may be taken as a one step more pessimistic approach when compared with the simple principle of maximum uncertainty consisting in the substitution of $\boldsymbol{x}^+$ for each $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$.

Let us introduce some examples when the idea beyond the mapping $\varphi(\mathcal{S})$ may become perhaps more intuitive.

(i) Let $\mathcal{S} = \emptyset$. Then, for each $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$ and no matter which the set $F(\boldsymbol{x}) = \{i \in \mathcal{N} : (\boldsymbol{x})_i = \lambda\}$ may be, $F(\boldsymbol{x})$ cannot be covered by a set from $\mathcal{S}$ simply because there are no sets $\mathcal{S}$ (it is the case also when $F(\boldsymbol{x}) = \emptyset$). Hence, $(\varphi(\mathcal{S}))(\boldsymbol{x}) = 1^\infty$ for every $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$ and $I((\varphi(\mathcal{S}))\pi) = 1^\infty$ (let us recall that I denotes the lattice-valued possibilistic entropy function defined by (2.4)).

(ii) Let $\mathcal{S} = \{\emptyset\}$. Then $F(\boldsymbol{x}) \subset S \in \mathcal{S}$ is the case iff $F(\boldsymbol{x}) = \emptyset$, hence, iff $\boldsymbol{x} \in \{0,1\}^\infty$ holds. So, $(\varphi(\mathcal{S}))(\boldsymbol{x}) = \boldsymbol{x}$, if $\boldsymbol{x} \in \{0,1\}^\infty$, $(\varphi(\mathcal{S}))(\boldsymbol{x}) = 1^\infty$ otherwise. Consequently, if Ω contains at least two elements and if both the sequences $\pi(\omega_1), \pi(\omega_2)$ contain at least one occurrence of λ, then $I((\varphi(\mathcal{S}))\pi) = 1^\infty$.

(iii) Let $\mathcal{S} = \{\mathcal{N}\}, \mathcal{N} = \{1,2\ldots\}$. Then, for each $\boldsymbol{x} = \{0,1,\lambda\}^\infty$, $F(\boldsymbol{x}) \subset \mathcal{N} \in \mathcal{S}$ holds, so that $(\varphi(\mathcal{S}))(\boldsymbol{x}) = \boldsymbol{x}^+$, i.e., all occurrences of λ in $\boldsymbol{x}$ are replaced by 1, leaving the 0- and 1-values untouched. This is the most simple maximization of the entropy function I as briefly re-called in Section 2.

(iv) Let $\mathcal{S} = \{A\}, \emptyset \neq A \subset \mathcal{N}, A \neq \mathcal{N}$. Then $(\varphi(\mathcal{S}))(\boldsymbol{x}) = \boldsymbol{x}^+$, if each i such that $(\boldsymbol{x})_i = \lambda$ belongs to A, $(\varphi(\mathcal{S}))(\boldsymbol{x}) = 1^\infty$ otherwise. Hence, only the values with dimensions in A may be approximated by the local application of the maximum uncertainty principle. If some value $(\boldsymbol{x})_i$ with i outside A is missing, all the sequence $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$ is doubtful and 1^∞ is its only sure upper bound.

(v) Let $R \in \mathcal{N}$, let $\mathcal{S} = \{A \subset \mathcal{N} : \|A\| \leq R\}$, $\|A\|$ stands for the cardinality of A. Then only the sequences $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$ where the number of unknown values does not exceed R can be approximated locally, i.e., by $\boldsymbol{x}^+$, setting $(\varphi(\mathcal{S}))(\boldsymbol{x}) = 1^\infty$ otherwise. Let us note explicitly that our model of approximation of $\boldsymbol{x}$ does not admit a repeated application of the operation $\varphi(\mathcal{S})$ to various R-tuples of indices from $\mathcal{N}$.

(vi) Let $\mathcal{S} = \mathcal{P}_f(\mathcal{N}) = \{A \subset \mathcal{N} : \|A\| < \infty\}$. Then only the sequences $x \in \{0,1,\lambda\}^\infty$ with a finite number of occurrences of λ can be approximated locally, i.e., $(\varphi(\mathcal{S}))(\boldsymbol{x}) = \boldsymbol{x}^+$. Let us note that the case $\mathcal{S} = \mathcal{P}_f(\mathcal{N})$ differs from (iii), which is equivalent to $\mathcal{S} = \mathcal{P}(\mathcal{N})$.

In what follows, the mapping $\varphi(\mathcal{S}) : \{0,1,\lambda\}^\infty \to \{0,1\}^\infty$ will be called the $\mathcal{S}$-*embedding of* $\{0,1,\lambda\}^\infty$ *into* $\{0,1\}^\infty$.

4 Some Results on $\mathcal{S}$-Embeddings

Given $\mathcal{S} \subset \mathcal{P}(\mathcal{N})$ and $\boldsymbol{x} = \{0,1,\lambda\}^\infty$, denote by $F(\boldsymbol{x}) \subset \{0,1\}^\infty$ the following set of binary sequences. If there exists $S \in \mathcal{S}$ such that $F(\boldsymbol{x}) \subset S$ holds, then

$$F(\boldsymbol{x}) = \{\boldsymbol{y} \in \{0,1\}^\infty : (\boldsymbol{y})_i = (\boldsymbol{x})_i \text{ for each } i \in \mathcal{N} \text{ such that } (\boldsymbol{x})_i \neq \lambda\}, \quad (4.1)$$

$G(\boldsymbol{x}) = \{0,1\}^\infty$ otherwise, i.e., if $F(\boldsymbol{x})$ cannot be covered by some $S \in \mathcal{S}$. When transforming $\boldsymbol{x}$ into $(\varphi(\mathcal{S}))(\boldsymbol{x})$ we observe easily that, for each $i \in \mathcal{N}$, either $((\varphi(\mathcal{S})(\boldsymbol{x}))_i = (\boldsymbol{x})_i$, or $((\varphi(\mathcal{S}))(\boldsymbol{x}))_i = 1$. Hence, for each $\mathcal{S} \subset \mathcal{P}(\mathcal{N})$, each $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$, and each $\boldsymbol{y} \in G(\boldsymbol{x})$ the inequality $\boldsymbol{y} \leq_T (\varphi(\mathcal{S}))(\boldsymbol{x})$ holds, the equality being the case iff $\boldsymbol{x} \in \{0,1\}^\infty$ holds. Moreover, for the same $\mathcal{S}$ and $\boldsymbol{x}$, if $(\boldsymbol{x})_i = \lambda$, then $((\varphi(\mathcal{S})(\boldsymbol{x}))_i = 1$ holds, so that $\boldsymbol{x}^+ \leq_0 (\varphi(\mathcal{S}))(\boldsymbol{x})$ follows, here $\boldsymbol{x}^+$ results from $\boldsymbol{x}$ when replacing all occurrences of λ in $\boldsymbol{x}$ by 1. However, as shown above (Section 3, (iii)), the binary sequence $\boldsymbol{x}^+$ can be also defined by $(\varphi(\{\mathcal{N}\}))(\boldsymbol{x})$, so that we arrive at the conclusion that $(\varphi(\{\mathcal{N}\}))(\boldsymbol{x}) \leq_0 (\varphi(\mathcal{S}))(\boldsymbol{x})$ holds for each $\mathcal{S} \subset \mathcal{P}(\mathcal{N})$ and each $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$. The possibilistic entropy function I, defined by (2.4), is defined only for binary sequences, so that we cannot directly define $I(\pi^T)$ for $\{0,1,\lambda\}^\infty$-valued possibilistic distribution π^T and compare it with the entropy value

$$I((\varphi(\mathcal{S}))(\pi)) = I(\{(\varphi(\mathcal{S}))(\pi(\omega)) : \omega \in \Omega\}). \quad (4.2)$$

So, we will compare the values $I((\varphi(\mathcal{S}))(\pi))$ and $I((\varphi(\{\mathcal{N}\}))(\pi))$.

Theorem 4.1 *Let $\mathcal{S} \subset \mathcal{P}(\mathcal{N})$, let π be a $\{0,1,\lambda\}^\infty$-valued possibilistic distribution on a nonempty set Ω, let I be the possibilistic entropy function defined by (2.4). Then the inequality*

$$I((\varphi(\{\mathcal{N}\}))(\pi)) \leq_0 I((\varphi(\mathcal{S}))(\pi)) \quad (4.3)$$

holds.

Proof: Both the systems $(\varphi(\{\mathcal{N}\}))(\pi) = \{(\varphi(\{\mathcal{N}\}))(\pi(\omega)) : \omega \in \Omega\}$ and $(\varphi(\mathcal{S}))(\pi) = \{(\varphi(\mathcal{S}))(\pi(\omega)) : \omega \in \Omega\}$ define $\{0,1\}^\infty$-valued possibilistic distributions on Ω and the relation $(\varphi(\{\mathcal{N}\}))(\pi(\omega)) \leq_0 (\varphi(\mathcal{S}))(\pi(\omega))$ holds for each $\omega \in \Omega$. Hence, denoting $(\varphi(\{\mathcal{N}\}))(\pi)$ by $\pi_1 : (\varphi(\mathcal{S}))(\pi)$ by π_2, and applying Lemma 2.1, we complete the proof of the assertion. □

Let $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{P}(\mathcal{N})$ be systems of subsets $\mathcal{N}$. We say that $\mathcal{S}_2$ *is covered* by $\mathcal{S}_1 (\mathcal{S}_2 \ll \mathcal{S}_1$, in symbols), if each set $S_2 \in \mathcal{S}_2$ is covered by some $S_1 \in \mathcal{S}_1$, i.e., if there exists, for each $S_2 \in \mathcal{S}_2$, a set $S_1 \in \mathcal{S}_1$ such that $S_2 \subset S_1$ holds. E.g., if $\mathcal{S}_1 = \{\mathcal{N}\}$, then $\mathcal{S}_2 \ll \mathcal{S}_1$ holds for each $\mathcal{S}_2 \subset \mathcal{P}(\mathcal{N})$, on the other side, if $\mathcal{S}_1 = \{\emptyset\}$, then $\mathcal{S}_2 \ll \mathcal{S}_1$ holds iff $\mathcal{S}_2 = \mathcal{S}_1$ or $\mathcal{S}_2 = \emptyset$. Theorem 4.1 can be generalized as follows.

Theorem 4.2 *Let $\mathcal{S}_1, \mathcal{S}_2$ be systems of subsets of $\mathcal{N}$ such that $\mathcal{S}_2$ is covered by $\mathcal{S}_1$, let π^T be a $\{0,1,\lambda\}^\infty$-valued possibilistic distribution on a nonempty set Ω, let I be the possibilistic entropy function defined by (2.4). Then the inequality*

$$I((\varphi(\mathcal{S}_1))(\pi)) \leq_0 I((\varphi(\mathcal{S}_2))(\pi)) \tag{4.4}$$

holds.

Proof: As in the proof of Theorem 4.1, we have to prove that the inequality $(\varphi(\mathcal{S}_1))(\pi(\omega)) \leq_0 (\varphi(\mathcal{S}_2))(\pi(\omega))$ holds for each $\omega \in \Omega_0$. As a matter of fact, the inequality $(\varphi(\mathcal{S}_1))(\boldsymbol{x}) \leq_0 (\varphi(\mathcal{S}_2))(\boldsymbol{x})$ holds for each $\boldsymbol{x} \in \{0,1,\lambda\}^\infty$.

Indeed, let $\boldsymbol{x} = \langle (\boldsymbol{x})_1, (\boldsymbol{x})_2, \ldots \rangle$, so that $(\boldsymbol{x})_i \in \{0,1,\lambda\}$ for each $i \in \mathcal{N}$. For each such i, if $F(\boldsymbol{x}) = \{i \in \mathcal{N} : (\boldsymbol{x})_i = \lambda\} \subset S_2$ and $i \in S_2$ holds for some $S_2 \in \mathcal{S}_2$, then $S_2 \subset S_1$ and $i \in S_1$ holds for some $S_1 \in \mathcal{S}_1$, so that $((\varphi(\mathcal{S}_2))(\boldsymbol{x}))_i = ((\varphi(\mathcal{S}_1))(\boldsymbol{x}))_i$ $(= (\boldsymbol{x})_i$, if $(\boldsymbol{x})_i \neq \lambda)$, or $= 1$, if $(\boldsymbol{x})_i = \lambda$ is the case.

If $F(\boldsymbol{x})$ cannot be covered by some $S_1 \in \mathcal{S}_1$, then $F(\boldsymbol{x})$ cannot be covered by no matter which $S_2 \in \mathcal{S}_2$, so that $(\varphi(\mathcal{S}_1))(\boldsymbol{x}) = (\varphi(\mathcal{S}_2))(\boldsymbol{x}) = 1^\infty$, hence, $((\varphi(\mathcal{S}_1))(\boldsymbol{x}))_i = ((\varphi(\mathcal{S}_2))(\boldsymbol{x}))_i = 1$ for each $i \in \mathcal{N}$ trivially follows. The case that $i \in S_2, F(\boldsymbol{x}) \subset S_2$, holds for some $S_2 \in \mathcal{S}_2$, but for no $S_1 \in \mathcal{S}_1$ the relation $i \in S_1, F(\boldsymbol{x}) \subset S_1$ is valid, is excluded by the assumption that $\mathcal{S}_2 \ll \mathcal{S}_1$ holds.

The only case for an index $i \in \mathcal{N}$ which remains to be analyzed reads that there is $S_1 \in \mathcal{S}_1$ such that $F(\boldsymbol{x}) \subset S_1$ and $i \in S_1$ holds, but there is no $S_2 \in \mathcal{S}_2$ with the property that $F(\boldsymbol{x}) \subset S_2$ and $i \in S_2$ holds together. In this case, however, $(\varphi(\mathcal{S}_2))(\boldsymbol{x}) = 1^\infty$, hence, $((\varphi(\mathcal{S}_2))(\boldsymbol{x}))_i = 1$, so that the inequality $((\varphi(\mathcal{S}_1))(\boldsymbol{x}))_i \leq ((\varphi(\mathcal{S}_2))(\boldsymbol{x}))_i$ holds again. To conclude, the last inequality holds for each $i \in \mathcal{N}$, so that the inequality $(\varphi(\mathcal{S}_1))(\pi(\omega)) \leq_0 (\varphi(\mathcal{S}_2))(\pi(\omega))$ holds for each $\omega \in \Omega$, hence, the assertion (4.4) follows. □

When introducing the relation $\mathcal{S}_2 \ll \mathcal{S}_1$, we oriented the inequalities $\ll$ in the way keeping the consistence with the increasing value of the entropy $I((\varphi(\mathcal{S}))(\pi))$. In what follows, we will need also the inverse relation $\ll_R$ defined, for each $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{P}(\mathcal{N})$, by the relation $\mathcal{S}_2 \ll_R \mathcal{S}_1$ iff $\mathcal{S}_1 \ll \mathcal{S}_2$ is the case. Both the relations $\ll$ and $\ll_R$ obviously define partial ordering on $\mathcal{P}(\mathcal{N})$. So, for each $\mathcal{S}_1 \subset \mathcal{P}(\mathcal{N})$ and for $\mathcal{S}_2 = \{\mathcal{N}\}$ the relation $\mathcal{S}_1 \ll_R \mathcal{S}_2$ is valid.

Definition 4.1 *A sequence $\{\mathcal{S}_n\}_{n=1}^\infty$ of subsets of $\mathcal{P}(\mathcal{N})$ tends to cover a set $A \subset \mathcal{N}$ ($\{\mathcal{S}_n\}_{n=1}^\infty \to_c A$, in symbols), if $\mathcal{S}_n \ll_R \mathcal{S}_{n+1}$ holds for each $n = 1, 2, \ldots$, and if there exists, for each finite $B \subset A, B \neq A$, an index $n_0 \in \mathcal{N}$ and a set $S_{n_0} \in \mathcal{S}_{n_0}$ such that $A \subset S_{n_0}, B \neq S_{n_0}$ holds.*

Obviously, if A is a finite subset of $\mathcal{N}$, then $\{\mathcal{S}_n\}_{n=1}^\infty \to_c A$ is the case iff there exists $n_0 \in \mathcal{N}$ and $S_{n_0} \in \mathcal{S}_{n_0}$ such that $A \subset S_{n_0}$ holds. On the other side, if A is infinite, say, if $A = \mathcal{N}$, then $\mathcal{N}$ itself need not be a member of any $\mathcal{S}_n$. E.g., if $\mathcal{S}_n = \{\{1, 2, \ldots, n\}\}_{n=1}^\infty \subset \mathcal{P}(\mathcal{N})$, then $\{\mathcal{S}_n\}_{n=1}^\infty \to_c \mathcal{N}$ holds.

A sequence $\{\boldsymbol{x}^i\}_{i=1}^{\infty}, \boldsymbol{x}^i \in \{0,1\}^{\infty}$ for each $i \in \mathcal{N}$, *tends to* $\boldsymbol{x}^0 \in \{0,1\}^{\infty}$, $\{\boldsymbol{x}^i\}_{i=1}^{\infty} \to \boldsymbol{x}^0$, in symbols, if there exists, for each $K \in \mathcal{N}$, an index $n_K \in \mathcal{N}$ such that, for each $n \geq n_K$ and each $j \leq K$, the relation $(\boldsymbol{x}^n)_j = (\boldsymbol{x}^0)_j$ holds. Hence, for each $K \in \mathcal{N}$, the initial segments of the length K for $\boldsymbol{x}^j$ and $\boldsymbol{x}^0$ are identical for j large enough.

Lemma 4.1 *Let $\boldsymbol{x} \in \{0,1,\lambda\}^{\infty}$ be a ternary sequence, let the set $F(\boldsymbol{x}) = \{i \in \mathcal{N} : (\boldsymbol{x})_i = \lambda\}$ be finite, let $\{\mathcal{S}_n\}_{n=1}^{\infty}$ be a sequence of subsets of $\mathcal{P}(\mathcal{N})$ which tends to cover the set $F(\boldsymbol{x})$. Then the sequence $\{(\varphi(\mathcal{S}_n))(\boldsymbol{x})\}_{n=1}^{\infty}$ of binary sequences tends to $\boldsymbol{x}^+$ in the sense that $(\varphi(\mathcal{S}_n))(\boldsymbol{x}) = \boldsymbol{x}^+$ for each $n \geq n_0$ for some $n_0 \in \mathcal{N}$.*

Proof: If $F(\boldsymbol{x})$ is finite, there exists $n_0 \in \mathcal{N}$ and $S_{n_0} \in \mathcal{S}_{n_0}$ such that $F(\boldsymbol{x}) \subset S_{n_0}$ holds. Hence, according to the definition of binary sequence $(\varphi(\mathcal{S}_{n_0}))(\boldsymbol{x})$, we obtain that $((\varphi(\mathcal{S}_{n_0}))(\boldsymbol{x}))_i = (\boldsymbol{x})_i \in \{0,1\}$, if $i \in \mathcal{N} - S_{n_0}$ holds. For $i \in S_{n_0}$, we obtain that $((\varphi(\mathcal{S}_{n_0}))(\boldsymbol{x}))_i = (\boldsymbol{x})_i$, if $(\boldsymbol{x})_i \neq \lambda$, and $((\varphi(\mathcal{S}_{n_0}))(\boldsymbol{x}))_i = 1$, if $(\boldsymbol{x})_i = \lambda$. As may be easily observed, $(\varphi(\mathcal{S}_{n_0}))(\boldsymbol{x}) = \boldsymbol{x}^+$ follows (let us recall that $\boldsymbol{x}^+$ results from $\boldsymbol{x}$ when replacing all occurrences of λ in $\boldsymbol{x}$ by 1). As $\mathcal{S}_n \ll \mathcal{S}_{n+1}$ holds, in each $\mathcal{S}_n$ with $n \geq n_0$ there exists a set S_n covering $F(\boldsymbol{x})$, so that $(\varphi(\mathcal{S}_n))(\boldsymbol{x}) = \boldsymbol{x}^+$ holds for each $n \geq n_0$. Consequently, $(\varphi(\mathcal{S}_n))(\boldsymbol{x})$ tends to $\boldsymbol{x}^+$ in the sense described in Lemma 4.1. □

Theorem 4.3 *Let $\boldsymbol{x} = \langle(\boldsymbol{x})_1, (\boldsymbol{x})_2, \ldots\rangle \in \{0,1,\lambda\}^{\infty}$, let $n \in \mathcal{N}$, let $\boldsymbol{x}^{[n]} \in \{0,1,\lambda\}^{\infty}$ be defined in this way: $(\boldsymbol{x}^{[n]})_i = (\boldsymbol{x})_i$, if $i \leq n$, $(\boldsymbol{x}^{[n]})_i = 1$, if $i > n$ holds. Let the set $F(\boldsymbol{x})$ be infinite, let a sequence $\{\mathcal{S}_n\}_{n=1}^{\infty}$ of subsets of $\mathcal{P}(\mathcal{N})$ tend to cover the set $F(\boldsymbol{x})$. Then there exists a sequence $\langle k(1), k(2), \ldots\rangle$ of positive integers such that the sequence $\{(\varphi(\mathcal{S}_{k(n)}))(\boldsymbol{x}^{[n]})\}_{n=1}^{\infty}$ of binary sequences tends to $\boldsymbol{x}^+$.*

Proof: Obviously, the sequence $\{\boldsymbol{x}^{[n]}\}_{n=1}^{\infty}$ tends to $\boldsymbol{x}$ for $n \to \infty$, and for each $n \in \mathcal{N}$ the relations $F(\boldsymbol{x}^{[n]}) = F(\boldsymbol{x}) \cap \{1, 2, \ldots, n\}$ and $F(\boldsymbol{x}) = \bigcup_{n=1}^{\infty} F(\boldsymbol{x}^{[n]})$ are valid. As $S(\boldsymbol{x})$ is infinite and each $F(\boldsymbol{x}^{[n]})$ is finite, i.e., each $F(\boldsymbol{x}^{[n]})$ is a proper subset of $F(\boldsymbol{x})$, there exists, for each $n \in \mathcal{N}$, an index $k(n) \in \mathcal{N}$ and a set $S_{k(n)} \in \mathcal{S}_{k(n)}$ such that $F(\boldsymbol{x}^{[n]}) \subset S_{k(n)}$ holds. The conditions imposed on $\{\mathcal{S}_j\}_{j=1}^{\infty}$ yield that $S_{k(n)} \subset S_{k(n)+1} \in \mathcal{S}_{k(n)+1}$ holds for some $S_{k(n)+1} \subset \mathcal{N}$. So, for each $n \in \mathcal{N}$ there exists $k(n) \in \mathcal{N}$ such that, for each $j \geq k(n)$, there exists $S_j \in \mathcal{S}_j$ with the property $F(\boldsymbol{x}^{[n]}) \subset S_j$ being valid. According to the way in which the sequence $(\varphi(\mathcal{S}_j))(\boldsymbol{x}^{[n]}), j \geq k(n)$, is defined, we obtain that $((\varphi(\mathcal{S}_j))(\boldsymbol{x}^{[n]}))_l = (\boldsymbol{x}^+)_e$ for each $l \leq n$. So, the binary sequences $(\varphi)\mathcal{S}_{k(n)}))(\boldsymbol{x}^{[n]})$ tend to $\boldsymbol{x}^+$ for $n \to \infty$.

According to what we have just proved, if $\langle k(1), k(2), \ldots\rangle$ is a sequence of positive integers meeting the conditions of the assertion of Theorem 4.3, each sequence $\langle m(1), m(2), \ldots\rangle$ of positive integers such that $k(n) \leq m(n)$ holds for each $n \in \mathcal{N}$, also meets the conditions of the assertion and, for each $n, s \in \mathcal{N}$, the relation $((\varphi(\mathcal{S}_{k(n)}))(\boldsymbol{x}^{[n]}))_s = ((\varphi(\mathcal{S}_{m(n)}))(\boldsymbol{x}^{[n]}))_s$ is valid (for $s > n$ this identity holds trivially, so that in this case $(\boldsymbol{x}^{[n]})_s = 1$ and this value is untouched by no matter which $\mathcal{S}$-projection from $\{0,1,\lambda\}^{\infty}$ into $\{0,1\}^{\infty}$). □

5 Conclusions

Having preferred the idea to present the introduced results together with their more or less detailed proofs, and having been obliged to keep the constraints

imposed on the extent of the contributions necessary for proceedings of scientific meetings like this one, the only solution which remains is to select some from the achieved results, leaving aside the other, perhaps also promising and interesting ones.

What should be introduced and analyzed in more detail, in a future work, is to replace the Boolean-like ordering $\leq_T$ on $|0,1,\lambda|^\infty$ (denoted by $\leq_0$ on $\{0,1\}^\infty$) by the lexicographical ordering $\leq_T$ on $\{0,1,\lambda\}^\infty$. Namely, for each $\boldsymbol{x},\boldsymbol{y} \in \{0,1,\lambda\}^\infty$ we define $\boldsymbol{x} \leq_L \boldsymbol{y}$, if $\boldsymbol{x} = \boldsymbol{y}$, or if $(\boldsymbol{x})_{i_0} < (\boldsymbol{y})_{i_0}$ holds for $i_0 = \min\{k \in \mathcal{N} : (\boldsymbol{x})_k \neq (\boldsymbol{y})_k\}$. The relation $\leq_L$ defines a linear ordering on $\{0,1,\lambda\}^\infty$, $\langle\{0,1,\lambda\}^\infty, \leq_L\rangle$ is a complete lattice, and, for each $\boldsymbol{x},\boldsymbol{y} \in \{0,1,\lambda\}^\infty$, if $\boldsymbol{x} \leq_T \boldsymbol{y}$ is the case, then $\boldsymbol{x} \leq_L \boldsymbol{y}$ holds as well. Let us define lexicographical entropy function I^L by (2.4), just with the partial ordering $\leq_T$ and induced operations $\bigvee^T$ and $\bigwedge_T$ replaced by $\leq_L, \bigvee^L$, and $\bigwedge_L$. Most results proved in this paper and dealing with the principle of maximum entropy value I remain to hold also for I^L when replacing $\leq_T$ by $\leq_L$ during our considerations over the values from $\{0,1,\lambda\}^\infty$.

Another way of improving the ideas and results presented in this work consists in enriching the structure of "uncertain" values, possibly taken by the members $(\boldsymbol{x})_i$ of sequences from $\{0,1,\lambda\}^\infty$, by more values. E.g., we may consider a set $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_K < 1$ of real numbers and replace the space $\{0,1,\lambda\}^\infty$ by the space $\{0,\lambda_1,\lambda_2,\ldots,\lambda_K,1\}^\infty$. Also this way of further research seems to be interesting and promising.

Let us hope that we will be able to develop, in more detail, at least some of these ideas in a future paper oriented towards the relevant field of research.

Acknowledgements

This work was partially supported by grant No. IAA100300503 of GA AS CR and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

References

[1] G. Birkhoff (1967). Lattice Theory, 3rd edition. Providence, Rhode Island.

[2] G. DeCooman (1997). Possibility theory I, II, III. International Journal of General Systems 25, pp. 291-323, 325-351, 353-371.

[3] D. Dubois, H. Prade (Eds.) (2000). Possibility theory, probability theory and fuzzy sets: misunderstandings, bridges and gaps. In: The Handbook of Fuzzy Sets Series – Fundamental of Fuzzy Sets. Kluwer Acadenic Publishers, Boston, pp. 343-438.

[4] R. Faure, E. Heurgon (1971). Structures Ordonnées et Algèbres de Boole. Gauthier-Villars, Paris.

[5] J. A. Goguen (1967). $\mathcal{L}$-fuzzy sets. Journal of Mathematical Analysis and Applications 18, pp. 145-174.

[6] G. J. Klir, M. Wierman (1999). Uncertainty-Based Information. Physica Verlag, Heidelberg, New York.

[7] I. Kramosil (2008). Locally sensitive lattice-valued possibilistic entropy functions. Neural Network World 18, pp. 469-488.

[8] I. Kramosil (2008). Lattice-valued possibilistic entropy measure. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 16, pp. 829-846.

[9] I. Kramosil (2008). Possibilistic distributions taking values in an incomplete lattice. Technical Report no V-1044, 7 p., Institute of Computer Science, Prague.

[10] D. Morales, L. Pardo, I. Vajda (1996). Uncertainty of discrete stochastic systems: general theory and statistical inference. IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans 26, pp. 681-697.

[11] R. Sikorski (1964). Boolean Algebras, 2nd edition. Springer, Berlin.

[12] I. Vajda (1989). Theory of Statistical Inference and Information. Kluwer Academic Publishers, Dordrecht.

[13] L. A. Zadeh (1978). Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, pp. 3-28.

Equivalence Problem in Compositional Models

Václav Kratochvíl
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
velorex@utia.cz

Abstract

Structure of each Compositional model can be visualized by a tool called persegram. Every persegram over a finite non-empty set of variables N induces an independence model over N, which is a list of conditional independence statements over N. The *equivalence problem* is how to characterize (in graphical terms) whether all independence statements in the model induced by persegram $\mathcal{P}$ are in the model induced by a second persegram $\mathcal{P}'$ and vice versa. In the previous paper [5] indirect characterization of equivalence was done. We introduced three different operations on persegrams remaining independence model which combined together are able to generate the (whole) class of equivalent persegrams. That characterization is indirect in the following sense: Two persegrams $\mathcal{P}$, $\mathcal{P}'$ are equivalent if there exists a sequence of persegrams from $\mathcal{P}$ to $\mathcal{P}'$ such that only so called IE-operations are performed to get next persegram in the sequence.

In this paper we give the motivation and introduction for direct characterization of equivalence. We have found some invariants among equivalent persegrams that have to be remained. In spite of that, the final simple direct characterization is not given. Instead we give several properties of equivalent persegrams that could be helpful.

1 Introduction

The ability to represent and process multidimensional probability distributions is a necessary condition for the application of probabilistic methods in Artificial Intelligence. Among the most popular approaches are the methods based on Graphical Markov Models, e.g., Bayesian Networks. The Compositional models are an alternative approach to Graphical Markov Models. These models are generated by a sequence (generating sequence) of low-dimensional distributions, which, composed together, create a distribution - the so called *Compositional model*. Moreover, while a model is composed together, a system of (un)conditional independencies is simultaneously introduced by the structure of the generating sequence.

The structure can be visualized by a tool called *persegram* and one can read induced independencies directly using this tool. That is why we can say that every persegram over a finite non-empty set of variables N induces an

independence model over N - a list of conditional independence statements over N. The *equivalence problem* is how to characterize (in graphical terms) whether all independence statements in the model induced by persegram $\mathcal{P}$ are also in the independence model induced by a second persegram $\mathcal{P}'$ and vice versa.

2 Compositional Models

A Bayesian network may be defined as a multidimensional distribution factorizing with respect to an acyclic directed graph. Alternatively, it may be defined by its graph and an appropriate system of low-dimensional (oligodimensional) *conditional distributions*. Contrary, Compositional models are defined as a multidimensional distribution assembled from a sequence of oligodimensional *unconditional distributions*, with the help of operators of composition. The main advantage of both approaches lies in the fact that oligodimensional distributions could be easily stored in a computer memory. However, computing with a multidimensional distribution that is split into many pieces may be exceptionally complicated. The advantage of Compositional models in comparison with Bayesian networks consists in the fact that compositional models explicitly express some marginals, whose computation in a Bayesian network may be demanding. Compositional model is assembled ,in contrast to Bayesian network, from unconditional distributions.

2.1 Notation and Basic Properties

Throughout the paper the symbol N will denote a non-empty set of finite-valued *variables*. From the next chapter on, variables will be represented by markers of a persegram. All probability distributions of this variables will be denoted by Greek letters (usually π, κ); thus for $K \subset N$, we consider a distribution (a probability measure over K) $\pi(K)$ which is defined for variables K. When several distributions will be considered, we shall distinguish them by indices. For a probability distribution $\pi(K)$ and $U \subset K$ we will consider a *marginal distribution* $\pi(U)$.

The following conventions will be used throughout the paper. Given sets $K, L \subset N$ the juxtaposition KL will denote their union $K \cup L$. The following symbols will be reserved for special subsets of N: K, R, S. The symbol U, V, W, Z will be used for general subsets of N. The symbol $|U|$ will be used to denote the number of elements of a finite set U, that is, its *cardinality*. u, v, w, z denotes variables as well as singletons $\{x\}, \ldots$

Independence and dependence statements over N correspond to special *disjoint triples* over N. Thy sumbol $\langle U, V|Z\rangle$ denotes a triplet of pirwise disjoint subsets U, V, Z of N. This notations anticipates the intended meaning: the set of variables U is conditionally independent or dependent of the set of variables V given the set of variables Z. This is why the third set Z is separated by a straight line: it has a special meaning of the conditioning set. The symbol $\mathcal{T}(N)$ will denote the class of all disjoint triplets over N:

$$\mathcal{T}(N) = \{\langle U, V|Z\rangle; U, V, Z \subseteq N \quad U \cap V = V \cap Z = Z \cap U = \emptyset\}$$

To describe how to compose low-dimensional distributions to get a distribution of a higher dimension we use the following operator of composition.

Definition 2.1. For arbitrary two distributions $\pi(K)$ and $\kappa(L)$ their *composition* is given by the formula

$$\pi(K) \rhd \kappa(L) = \begin{cases} \frac{\pi(K)\kappa(L)}{\kappa(K \cap L)} & \text{if } \pi^{\downarrow K \cap L} \ll \kappa^{\downarrow K \cap L}, \\ \text{undefined} & \text{otherwise,} \end{cases}$$

where the symbol $\pi(M) \ll \kappa(M)$ denotes that $\pi(M)$ is *dominated* by $\kappa(M)$, which means (in the considered finite setting)

$$\forall x \in \times_{j \in M} \mathbf{X}_j; (\kappa(x) = 0 \Longrightarrow \pi(x) = 0).$$

The result of the composition (if defined) is a new distribution. We can iteratively repeat the process of composition to obtain a multidimensional distribution - a model approximating the original distribution with corresponding marginals. That is why these multidimensional distributions (and the whole theory as well) are called *Compositional models*. To describe such a model it is sufficient to introduce an ordered system of low-dimensional distributions $\pi_1, \pi_2, \ldots, \pi_n$. If all compositions are defined, we call this ordered system a *generating sequence*. To get a distribution represented by this sequence one has to apply the operators from left to right:

$$\pi_1 \rhd \pi_2 \rhd \pi_3 \rhd \ldots \rhd \pi_{n-1} \rhd \pi_n = (\ldots((\pi_1 \rhd \pi_2) \rhd \pi_3) \rhd \ldots \rhd \pi_{n-1}) \rhd \pi_n.$$

From now on, we consider generating sequence $\pi_1(K_1), \pi_2(K_2), \ldots, \pi_n(K_n)$ which defines a distribution

$$\pi_1(K_1) \rhd \pi_2(K_2) \rhd \ldots \rhd \pi_n(K_n).$$

Therefore, whenever distribution π_i is used, we assume it is defined for variables K_i. In addition, each set K_i can be divided into two disjoint parts. We denote them R_i and S_i with the following sense:

$$R_i = K_i \backslash (K_1 \cup \ldots \cup K_{i-1}), S_i = K_i \cap (K_1 \cup \ldots \cup K_{i-1})$$

.

R_i denotes variables from K_i with the first appeared with respect to the sequence (meaning from left to right). S_i denotes the already used.

2.2 Graphical concepts

It is well-known that one can read conditional independence relations of a Bayesian network from its graph. A similar technique is used in compositional models. An appropriate tool for this is a *persegram*. Persegram is used to visualize the structure of a compositional model and is defined bellow.

Definition 2.2. Persegram $\mathcal{P}$ *of a generating sequence is a table in which rows correspond to variables (in an arbitrary order) and columns to low-dimensional distributions; ordering of the columns corresponds to the generating sequence ordering. A position in the table is marked if the respective distribution is defined for the corresponding variable. Markers for the first occurrence of each variable (i.e., the leftmost markers in rows) are squares (we call them* box-markers*) and for other occurrences there are* bullets.

Persegram $\mathcal{P}$ is a table of markers. Since the markers in the i-th column highlight variables for which generating sequence is defined, we denote markers in i-th column as K_i. Box-markers in i-th column of $\mathcal{P}$ are denoted like R_i and bullets like S_i. $K_i = R_i \cup S_i$. This notation is purposely in accordance with notation of variable sets in generating sequences to simplify readability and lucidity of the text.

Persegrams are usually denoted by $\mathcal{P}$ and if it is not specified otherwise $\mathcal{P}$ corresponds to the generating sequence $\pi_1(K_1), \ldots, \pi_n(K_n)$ where $K_1 \cup \ldots \cup K_n = N$. We say that $\mathcal{P}$ is defined over N. (i.e. $\mathcal{P}$ over N has n columns with markers $K_1, \ldots, K_n$ where $K_1 \cup \ldots \cup K_n = N$.)

To simplify the notation we will use the following symbol: Let $\mathcal{P}$ be a persegram over N. We introduce a function $]\![_{\mathcal{P}}\colon N \to \mathbb{N}$, which for every variable $u \in N$ returns the index of set K_i with the first appearance of u in the persegram $\mathcal{P}$. Due to the previously established notation can be said that $K_{]u[_{\mathcal{P}}}$ is a column K_i where $u \in R_i$. In other words: $]u[_{\mathcal{P}} = i : u \in R_i$.

Definition 2.3. *Let $\mathcal{P}$ be a persegram over N and $\preceq_{\mathcal{P}}$ a binary relation. For arbitrary $u, v \in N$ $u \preceq_{\mathcal{P}} v$ if $]u[_{\mathcal{P}} \leq]v[_{\mathcal{P}}$. Moreover we introduce the relation $\prec_{\mathcal{P}}$: $u \prec_{\mathcal{P}} v \Leftrightarrow u \preceq_{\mathcal{P}} v$ AND $v \not\preceq_{\mathcal{P}} u$.*

The following convention will be used throughout the paper: Given variables $u, v, w \in N$ and $\mathcal{P}$ over N, the term $u, v \prec_{\mathcal{P}} w$ denotes that $u \prec_{\mathcal{P}} w$ and $v \prec_{\mathcal{P}} w$. The symbol $\mathcal{P}$ may be omitted, if the content is clear.

2.3 Conditional independence

Conditional independence statements over N induced by the structure of Compositional model can be read from its persegram. Such independence is indicated by the absence of a *trail connecting or avoiding relevant markers*. It is defined below.

Definition 2.4. *Consider a persegram over N and a subset $Z \subset N$. A sequence of markers $m_0, \ldots, m_t$ is called a* Z-avoiding trail *that connects m_0 and m_t if it meets the following 4 conditions:*

1. *for each $s = 1, \ldots, t$ a couple (m_{s-1}, m_s) is in the same row (i.e., horizontal connection) or in the same column (vertical connection);*
2. *each vertical connection must be adjacent to a box-marker (one of the markers is a box-marker);*
3. *no horizontal connection corresponds to a variable from Z;*
4. *vertical and horizontal connections regularly alternate with the following possible exception: two vertical connections may be in direct succession if their common adjacent marker is a box-marker of a variable from Z;*

If a Z-avoiding trail connects two-box markers corresponding to variables u and v, we also say that these variables are connected by a Z-avoiding trail. *Suppose $\langle U, V|Z\rangle \in \mathcal{T}(N)$ is a disjoint triplet over N. One says that U and V are conditionally dependent by Z, written $U \not\perp\!\!\!\perp V|Z[\mathcal{P}]$, if there exists a Z-avoiding trail between variable $u \in U$ and variable $v \in V$ in $\mathcal{P}$. In the opposite case one says that U and V are conditionally independent by Z in $\mathcal{P}$, written $U \perp\!\!\!\perp V|Z[\mathcal{P}]$.*

We also say that $\langle U, V|Z\rangle$ *is* represented *in* $\mathcal{P}$. *The induced independence model* $\mathcal{I}(\mathcal{P})$ *and the induced dependence model* $\mathcal{D}(\mathcal{P})$ *are defined as follows:*

$$\mathcal{I}_{\mathcal{P}} = \{\langle U, V|Z\rangle \in \mathcal{T}(N); U \perp\!\!\!\perp V|Z[\mathcal{P}]\}$$

$$\mathcal{D}_{\mathcal{P}} = \{\langle U, V|Z\rangle \in \mathcal{T}(N); U \not\perp\!\!\!\perp V|Z[\mathcal{P}]\}$$

Example 2.5. *Consider persegram from Figures 1 and 2.*

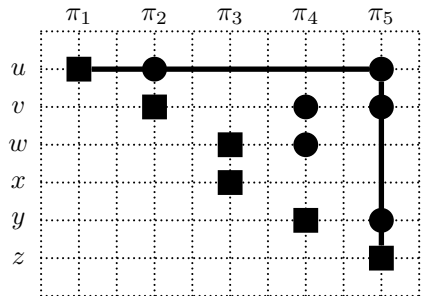


Figure 1: $\mathcal{P} : u \not\perp\!\!\!\perp z|\emptyset, u \not\perp\!\!\!\perp z|v$

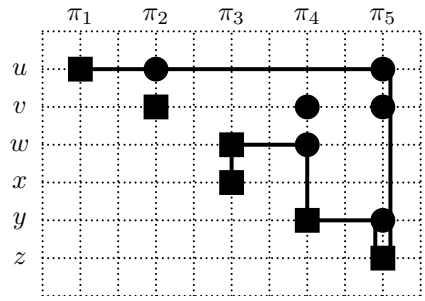


Figure 2: $\mathcal{P} : u \not\perp\!\!\!\perp x|z$

In Figure 1 a $\emptyset$*-avoiding trail is depicted. Therefore* $u \not\perp\!\!\!\perp z|\emptyset$. *Moreover, one can replace* $\emptyset$ *by any subset of* $\{v, w, x, y\}$ *which is avoiding as well. In Figure 2 is depicted another trail connecting* u *and* x *this time. Therefore* $u \not\perp\!\!\!\perp x|z$. *On the contrary to Figure 1 one can not replace* z *by any other variable except of* v. *Otherwise the condition 3. from the Definition 2.4 will be corrupted. (i.e.* $u \perp\!\!\!\perp x|y[\mathcal{P}]$ *for example)*

From the previous Definition 2.4 one can almost immediately get an interesting fact about variables appeared for the first time in the last column.

Lemma 2.6. *Consider a persegram* $\mathcal{P}$ *with* n *columns* $K_1, \ldots, K_n$ *and distinct variables* $u, v \in K_1 \cup \ldots \cup K_n$ *such that* $u \notin K_n$ *and* $v \in R_n$. *Then* $u \perp\!\!\!\perp v|S_n[\mathcal{P}]$.

Proof. Since v belongs to the last column of $\mathcal{P}$ only and u do not, every trail to v has to contain a horizontal connection to n-th column corresponding to some variable from S_n. By condition 3. of the Definition 2.4: No horizontal connection can correspond to variable from S_n. Then a S_n-avoiding trail between u and v can not exist. □

The following theorem shows an important parallel between independence read from compositional model and from its persegram. This theorem is given without proof, one can find it in [1].

Theorem 2.7. *Consider a generating sequence* $\pi_1(K_1), \ldots, \pi_n(K_n)$, *its corresponding persegram* $\mathcal{P}$, *and three disjoint subset* $U, V, Z \subset K_1 \cup \ldots \cup K_n$ *such that* $U \neq \emptyset \neq V$. *Then:*

$$U \perp\!\!\!\perp V|Z[\mathcal{P}] \Rightarrow U \perp\!\!\!\perp V|Z[\pi_1 \rhd \ldots \rhd \pi_n].$$

Notice that in definition 2.4 there is no condition concerning the order of rows in persegrams. This is not surprising because there is no rows ordering in definition 2.2 either.

To simplify proofs done by induction on the number of columns we introduce the concept of the *subpersegram* induced by subset of variables U. Unlike the subgraph which contains exactly those variables that induce it, subpersegram induced by a set U may be defined for some superset of U.

Definition 2.8. *Let $\mathcal{P}$ be a persegram over N. $U \subseteq N$. A* subpersegram *$\mathcal{P}[U]$ induced by U is the minimal left part of $\mathcal{P}$ containing all box-markers corresponding to U.*

Example 2.9. *Let $\mathcal{P}$ be the persegram represented in Example 2.5. Then the corresponding induced subpersegram $\mathcal{P}[z]$ is in Figure 3 and induced subpersegram $\mathcal{P}[w]$ is in Figure 4.*

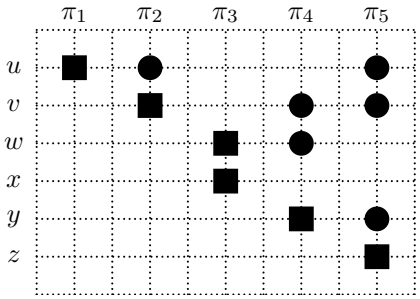


Figure 3: $\mathcal{P} \equiv \mathcal{P}[z]$

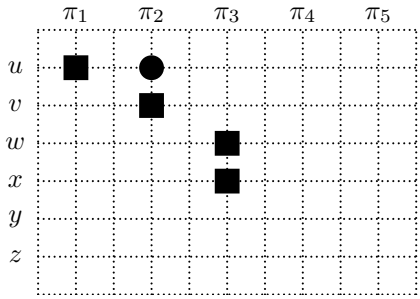


Figure 4: $\mathcal{P}[w]$

Lemma 2.10. *Let $\mathcal{P}$ be a persegram over N, and $u \not\perp\!\!\!\perp v | Z[\mathcal{P}]$. Then all Z-avoiding trails connecting u with v are in subpersegram $\mathcal{P}[u \cup v \cup Z]$ too.*

Proof. Suppose that $u \not\perp\!\!\!\perp v | Z[\mathcal{P}]$ and that there is a trail with a connection in $\mathcal{P}$ but not in $\mathcal{P}[u \cup v \cup Z]$. Let m is the first marker on the trail from u to v which belongs to such a column. Because this marker is the first one in such a column, a horizontal connection was used and therefore m is a bullet. Now one has to continue with a vertical connection(down) to box-marker. This box-marker does not correspond to any variable from Z (this column is not in $\mathcal{P}[u \cup v \cup Z]$). Therefore one has to continue with horizontal connection (to the right, this is a box-marker - there is nothing on left in the same row) to a bullet. Then down to a box-marker which does not correspond to any variable from Z etc. From such a trail is no return. Therefore such a trail can not exist. □

This lemma basically means, that if we are interested in relation $u \perp\!\!\!\perp v | Z[\mathcal{P}]$ we may focus on the subpersegram $\mathcal{P}[u \cup v \cup Z]$ only. This observation is summarized in the following corollary.

Corollary 2.11. *Let $\mathcal{P}$ be a persegram over N and $u, v \in N, Z \subset N \setminus \{u, v\}$. Then $u \perp\!\!\!\perp v | Z[\mathcal{P}[u \cup v \cup Z]] \Leftrightarrow u \perp\!\!\!\perp v | Z[\mathcal{P}]$.*

The following specific notation for certain composite dependence statements will be useful. Given a persegram $\mathcal{P}$ over N, distinct variables $u, v \in N$ and disjoint set $U \subseteq N \setminus \{u, v\}$ the symbol $u \not\perp\!\!\!\perp v | + U[\mathcal{P}]$ will be interpreted as the condition

$$\forall W \text{ such that } U \subseteq W \subseteq N \setminus \{u, v\} \text{ one has } u \not\perp\!\!\!\perp v | W[\mathcal{P}].$$

In words, u and v are (conditionally) dependent in $\mathcal{P}$ given any superset of U. If U is empty we write $*$ instead of $+U$. In particular, the following two symbols will be sometimes used

$$u \not\perp\!\!\!\perp v| * [\mathcal{P}] \equiv \forall W \text{ such that } W \subseteq N \setminus \{u, v\} \; u \not\perp\!\!\!\perp v|W[\mathcal{P}].$$

for distinct nodes $u, v \in N$, and

$$u \not\perp\!\!\!\perp v| + w[\mathcal{P}] \equiv \forall W \text{ such that } \{w\} \subseteq W \subseteq N \setminus \{u, v\} \text{ one has } u \not\perp\!\!\!\perp v|W[\mathcal{P}].$$

for distinct nodes $u, v, w \in N$. We give a certain graphical characterization of composite dependence statements of this kind below.

3 Equivalence problem

By the equivalence problem we understand the problem how to recognize whether two given persegrams $\mathcal{P}_1, \mathcal{P}_2$ over N induce the same independence model ($\mathcal{I}_{\mathcal{P}_1} = \mathcal{I}_{\mathcal{P}_2}$). It is of special importance to have an easy rule to recognize that two persegrams are equivalent in this sense and an easy way to convert $\mathcal{P}_1$ into $\mathcal{P}_2$ in terms of some elementary operations on persegrams. Another very important aspect is the ability to generate all persegrams which are equivalent to a given persegram.

Definition 3.1. *Persegrams $\mathcal{P}_1, \mathcal{P}_2$ (over the same variable set N) are called* independence equivalent*, if they induce the same independence model $\mathcal{I}_{\mathcal{P}_1} = \mathcal{I}_{\mathcal{P}_2}$.*

Remark 3.2. *One may easily see that the above mentioned definition could be formulated with the term of dependence model.* Persegrams $\mathcal{P}_1, \mathcal{P}_2$ (over the same variable set N) are independence equivalent, iff $\mathcal{D}_{\mathcal{P}_1} = \mathcal{D}_{\mathcal{P}_2}$. *This alternative is used in most proofs primarily.*

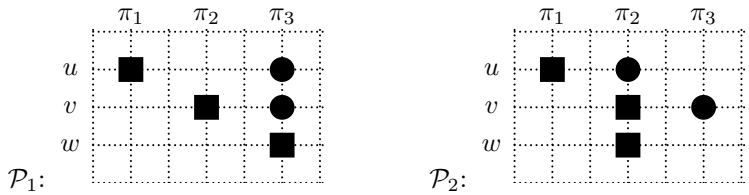
Like in Bayesian networks, it may happen that different persegrams induce the same independence model.

Example 3.3. *1. The following example is simple: $N = \{u, v\}$ and the following two persegrams $\mathcal{P}_1, \mathcal{P}_2$:*

$\mathcal{I}_{\mathcal{P}_1} = \mathcal{I}_{\mathcal{P}_2} = \{\langle u, v|\emptyset\rangle\}$ in this case.

2. On the other hand, the persegrams which have the same variable sets in columns in different order do not have to be equivalent. Let $N = \{u, v, w\}$ and consider the following persegrams:

$u \perp\!\!\!\perp v|\emptyset[\mathcal{P}_1]$ but $u \not\perp\!\!\!\perp v|\emptyset[\mathcal{P}_2]$. On the contrary, $u \not\perp\!\!\!\perp v|w[\mathcal{P}_1]$ and $u \not\perp\!\!\!\perp v|w[\mathcal{P}_2]$. The order of the columns in persegram is important.



3.1 Direct characterization

The solution of equivalence problem can be done in several ways. Some kind of *indirect characterization* of equivalence follows was done in the paper [5] where four special operations on persegrams were introduced. These operations are called *IE operations* (Independence equivalent) and they preserve independence statements induced by a persegram. These operations give us a tool to equivalence recognition: If two persegrams can be transformed from one to the other by a sequence of IE operations, then the persegrams are independence equivalent. Anyway, this characterization is indirect in the sense that, if two persegrams over same set of variables are given, then searching of such a sequence can be time demanding or even impossible. However, indirect characterization offers a method to generate a class of equivalent persegrams.

We are more interested in some type of *direct characterization* which allows us to decide on equivalence "immediately". This characterization should be based on some independence equivalence invariants.

Definition 3.4. *Let $\mathcal{P}$ be a persegram over N and $u, v \in N$ be two distinct variables. u, v are* connected *in $\mathcal{P}$ ($u \leftrightarrow v[\mathcal{P}]$) if there is a column in $\mathcal{P}$ containing markers of both variables and where at least one of them is a box-marker. Otherwise u, v are* disconnected *($u \nleftrightarrow v$).*

The following convention will be used thorough the paper: Given variables $u, v, w \in N$ and $\mathcal{P}$ over N, the term $u, v \leftrightarrow w$ denotes that $u \leftrightarrow w$ and $v \leftrightarrow w$.

For the purpose of the following text one should realize the obvious parallel between relation $u \leftrightarrow v$ and columns order and content. This parallel is summarized in the following remark.

Remark 3.5. *Let u, v are two different variables in $\mathcal{P}$ and $u \preceq_{\mathcal{P}} v$. Then $u \leftrightarrow v[\mathcal{P}] \Leftrightarrow u \in K_{]v[}$.*

Lemma 3.6. *Let $\mathcal{P}$ be a persegram over N and $u, v \in N$ are distinct variables, $u \preceq_{\mathcal{P}} v$. Then*

$$u \perp\!\!\!\perp v | S_{]v[}[\mathcal{P}] \Leftrightarrow u \nleftrightarrow v[\mathcal{P}]$$

Proof. $\Rightarrow$ Suppose $u \perp\!\!\!\perp v | S_{]v[}[\mathcal{P}]$ and $u \leftrightarrow v[\mathcal{P}]$. Since $u \preceq v$ then by Remark 3.5 $u \in S_{]v[}$. This however contradicts with the fact that sets involved in independence statements are not disjoint.

$\Leftarrow$ Suppose $u \nleftrightarrow v$ and $u \not\perp\!\!\!\perp v | S_{]v[}[\mathcal{P}]$. Since $u \preceq v$, and $S_{]v[} \prec v$ then by Lemma 2.6 $u \perp\!\!\!\perp v | S_{]v[}[\mathcal{P}[v]]$. By corollary 2.11 $u \perp\!\!\!\perp v | S_{]v[}[\mathcal{P}]$, which contradicts with assumptions.

□

Anyway, please realize that because of using an induced subpersegram $\mathcal{P}[v]$ in the proof, the equation $u \preceq v; u \nleftrightarrow v \Leftrightarrow u \perp\!\!\!\perp v| + S_{]v[}[\mathcal{P}]$ generally does not hold.

With the help of the previous lemma one can prove the following important assertion.

Lemma 3.7. *Let $\mathcal{P}$ be a persegram over N and $u, v \in N$ are distinct variables. Then*

$$u \leftrightarrow v[\mathcal{P}] \Leftrightarrow u \not\perp\!\!\!\perp v| * [\mathcal{P}].$$

Proof. $\Rightarrow$ Let $u \leftrightarrow v$ and $u \perp\!\!\!\perp v|w$ where $w \in N \setminus \{u, v\}$. Because $u \leftrightarrow v$ then the trail $u \rightsquigarrow_{\emptyset} v$ consists of one vertical and perhaps one horizontal connection and avoid any $w \in N \setminus \{u, v\}$. It contradicts the fact $a \perp\!\!\!\perp v|w$.

$\Leftarrow$ Suppose $u \not\perp\!\!\!\perp v|*$ and $u \nleftrightarrow v$; one can assume without loss of generality that $u \prec v$. Then $u \perp\!\!\!\perp v|S_{]v[}$ according to the lemma 3.6.

□

The previous two lemmata shows an interesting invariant of independence equivalence. Two persegrams, if equivalent, have the same set of connections.

Definition 3.8. *Let $\mathcal{P}$ be a persegram over N. A* connection set *$\mathcal{E}(\mathcal{P})$ is a set of all pairs $\langle u, v\rangle : u, v \in N$, where $u \leftrightarrow v[\mathcal{P}]$. $\mathcal{E}(\mathcal{P}) = \{\langle u, v\rangle : u, v \in N, u \leftrightarrow v[\mathcal{P}]\}$*

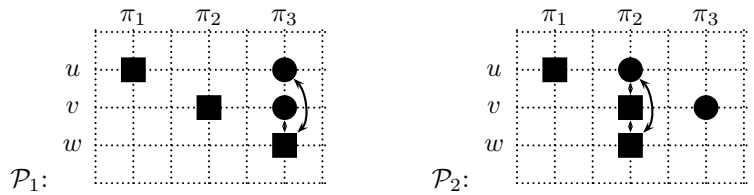
Corollary 3.9. *Let $\mathcal{P}, \mathcal{P}'$ are persegrams over N. If $\mathcal{I}_{\mathcal{P}} = \mathcal{I}_{\mathcal{P}'}$ then $\mathcal{E}(\mathcal{P}) = \mathcal{E}(\mathcal{P}')$.*

Example 3.10. *In the Example 3.3 four different persegrams are shown. The first two are equivalent, the second two are not. Let us show this example again with knowledge of the previous lemma.*

1. *Let $\mathcal{P}_1, \mathcal{P}_2$ are the following simple persegrams over $N = \{u, v\}$: One can*

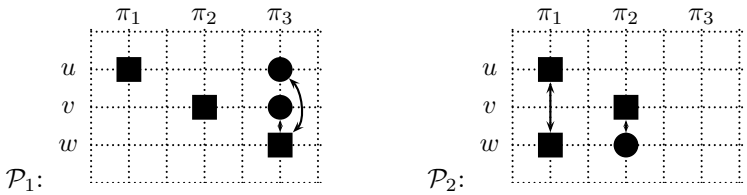
   easily see that $\mathcal{E}(\mathcal{P}_1) = \mathcal{E}(\mathcal{P}_2) = \emptyset$. The claim $\mathcal{I}_{\mathcal{P}_1} = \mathcal{I}_{\mathcal{P}_2} = \{\langle u, v|\emptyset\rangle\}$ is known from the Example 3.3.

2. *On the other hand, consider the following persegrams over $N = \{u, v, w\}$. Connections between variables are highlighted by arrows.*



   Thanks to Example 3.3 one knows that $\mathcal{I}_{\mathcal{P}_1} \neq \mathcal{I}_{\mathcal{P}_2}$. Since $\mathcal{E}(\mathcal{P}_1) = \{\langle u, w\rangle, \langle v, w\rangle\}$ but $\mathcal{E}(\mathcal{P}_2) = \mathcal{E}(\mathcal{P}_1) \cup \{\langle u, v\rangle\}$, non-equivalence is obvious now.

3. *Anyway, there exist persegrams* $\mathcal{P}_1, \mathcal{P}_2$ *where* $\mathcal{E}(\mathcal{P}_1) = \mathcal{E}(\mathcal{P}_2)$ *but* $\mathcal{I}_{\mathcal{P}_1} \neq \mathcal{I}_{\mathcal{P}_2}$.



$\mathcal{E}(\mathcal{P}_1) = \{\langle u, w\rangle, \langle v, w\rangle\} = \mathcal{E}(\mathcal{P}_2)$. *However* $\mathcal{I}_{\mathcal{P}_1} \neq \mathcal{I}_{\mathcal{P}_2}$ *since* $u \not\perp\!\!\!\perp v | w[\mathcal{P}_1]$ *and* $u \perp\!\!\!\perp v | w[\mathcal{P}_2]$.

It follows from the previous example, that the previous invariant is not strong enough to ensure the equivalence. It is necessary to try to find an another invariant.

When one consider a relation $\preceq_{\mathcal{P}}$, then every persegram satisfy some partial variables ordering. For example, $u \prec v \prec w$ in persegram $\mathcal{P}_1$ but $u \preceq w \prec v$ in persegram $\mathcal{P}_2$ in the third part of the previous Example 3.10. Is it possible that the order of the variables will be some kind of invariant? It will be definitely not in that simple way. It can be easily seen in the first part of the previous Example 3.10, where $u \prec v$ in $\mathcal{P}_1$ but $v \prec u$ in $\mathcal{P}_2$.

Two equivalent persegrams may have different ordering of variables. If, however, we are interested in the ordering of several specially connected variables only, then we obtain an another invariant of independence equivalence. It is based on *Ordering conditions* defined bellow.

Definition 3.11. *Let* $\mathcal{P}$ *be a persegram over* N. *An* Ordering condition *is a triplet of variables* $u, v, w \in N$ *where* $u, v \prec w$; $u, v \leftrightarrow w$; *and* $u \nleftrightarrow v$ *in* $\mathcal{P}$. *Such an ordering condition is denoted by* $[u, v] \prec w[\mathcal{P}]$.

An example of an ordering condition can be found it the second and third part of the Example 3.10 in $\mathcal{P}_1$. $[u, v] \prec w[\mathcal{P}_1]$ in that case. Persegrams $\mathcal{P}_2$ from both those parts of that Example do not contain any ordering condition.

Lemma 3.12. *Let* $\mathcal{P}$ *be a persegram over* N, $u, v, w \in N$ *distinct nodes. Then*

$$[u, v] \prec w \Leftrightarrow u \not\perp\!\!\!\perp v | + w[\mathcal{P}].$$

Proof. Suppose $[u, v] \prec w[\mathcal{P}]$. By Remark 3.5 $u, v \in S_n$. Then Let $W \subset N \setminus \{u, v\}$ such that $w \in W$. As one can see on the Figure 5 W-avoiding trail composed from u, w, v connects u, v for every W. Hence, $u \not\perp\!\!\!\perp v | W[\mathcal{P}]$ for every W (denoted by $u \not\perp\!\!\!\perp v | + w[\mathcal{P}]$). To prove sufficiency suppose by contradiction that $u \succ w \vee v \succ w$. (Since $u \nleftrightarrow v$, one does not have to consider equalities in $u \succeq w, v \succeq w$) Suppose $u \prec v$, then by remark 3.5 $w \in S_{]v[}$. Since $u \nleftrightarrow v$ the lemma 3.6 leads to contradiction. □

The above mentioned invariant can be easily concluded into the following implication.

Corollary 3.13. *Let* $\mathcal{P}, \mathcal{P}'$ *be two persegrams over* N. *If* $\mathcal{I}_{\mathcal{P}} = \mathcal{I}_{\mathcal{P}'}$ *then* $\mathcal{E}(\mathcal{P}) = \mathcal{E}(\mathcal{P}')$ *and they induce the same set of ordering conditions.*

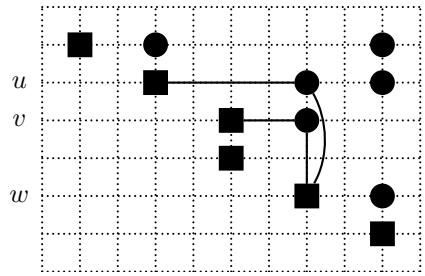


Figure 5: $u \not\perp\!\!\!\perp v| + w$

The question is: Does this implication hold also in the opposite direction? I.e. If two persegrams $\mathcal{P}, \mathcal{P}'$ over the same set N induce the same ordering conditions and $\mathcal{E}(\mathcal{P}) = \mathcal{E}(\mathcal{P}')$, are $\mathcal{P}, \mathcal{P}'$ independence equivalent? The answer for this question is still unknown. Despite the fact that all experiments confirm this theory, the formal proof has not been finished yet.

4 Conclusion

In this paper we gave a short introduction into equivalence problem. This problem includes several sub-problems where one of them is how to simply recognize whether two given persegrams are equivalent. One can say, how to recognize equivalence "on the first sight". The solution to this problem is a direct characterization involving some invariants sufficient for equivalence decision.

Two invariants we introduced: *Connections set* and *Ordering conditions*. Are these invariants sufficient to decide whether two given persegrams are equivalent? This question remains open.

References

[1] R. Jiroušek. *Multidimensional Compositional Models*. Preprint DAR - ÚTIA 2006/4, ÚTIA AV ČR, Prague, (2006).

[2] T. Kočka, R. R. Bouckaert, M. Studený. *On the Inclusion Problem*. Research report 2010, ÚTIA AV ČR, Prague (2001).

[3] M. Studený. *O strukturách podmíněné nezávislosti*. Rukopis série přednášek. Prague (2008).

[4] R. Merris: Graph Theory. Wiley Interscience, New York 2001.

[5] V. Kratochvíl. *Equivalence Problem in Compositional Models*. Doktorandské dny 2008, Nakladatelství ČVUT, Praha, p.125-134, 2008.

Selecting Marginals for Decision-Making Based on Marginal Problem

Otakar Kříž

Prague
o.kriz@upcmail.cz

Abstract

Diagnostic problem consists in finding a value of a *diagnostic variable* η on the basis of concrete values of some *symptom variables* $\xi_1, \xi_2, \cdots \xi_n$. The link between diagnosis and symptoms is supposed not to be a strict functional dependence, but there is certain uncertainty involved. One of the theoretical approaches to this decision making under uncertainty is based on the so called *marginal problem*. The "knowledge base" for an inference engine (i.e. algorithm) that performs this decision making is formed by a set of marginals. Given certain conditions, the paper suggests an heuristic algorithm for selecting marginals for which inference engines achieve the best decision making.

1 Introduction

The layout of the paper is the following one: Basic notions are introduced in Section 1, then goes the description of the algorithm in Section 2, experimental results in Section 3 and concluding remarks and recommendations in Section 4. [2ex]

1.1 Basic Setting

Let us suppose $(\Omega, \mathcal{X}, P)$ is a probabilistic space on which random variables η, $\xi_1, \xi_2, \cdots \xi_n$ are defined. *Diagnostic variable* η takes its values in a finite set of diagnoses $\{d_j\} = \mathrm{R}(\eta)$. (Symbol $\mathrm{R}(\vartheta)$ applied on a variable ϑ will denote its range (or codomain) in the sequel.) It is assumed the aim of the decision making is finding the most probable value of the η. All other variables, taking their values from finite sets denoted as $\mathrm{R}(\xi_1)$, $\mathrm{R}(\xi_2) \cdots$, $\mathrm{R}(\xi_n)$ are called *symptom variables* since their known values represent symptoms from which the unknown final diagnosis is inferred during decision making. (Sometimes, denotation Ξ will be used for set of all symptom variables.) Then, the set of all possible combinations of values of variables η, $\xi_1, \xi_2, \cdots \xi_n$ (i.e. their sample space), denoted as $\mathrm{R}(\eta, \xi_1, \xi_2, \cdots \xi_n)$, is a cartesian product of respective codomains:

$$\mathrm{R}\left(\eta, \xi_1, \xi_2, \cdots \xi_n\right) = \mathrm{R}(\eta) \times \mathrm{R}(\xi_1) \times \mathrm{R}(\xi_2) \cdots \mathrm{R}(\xi_n)$$

The mutual "behaviour" of $\eta, \xi_1, \xi_2, \cdots \xi_n$ is described completely by the joint distribution $P_{\eta\xi_1\xi_2\ldots\xi_n}$ induced from P and defined on $\mathrm{R}(\eta, \xi_1, \xi_2, \cdots \xi_n)$.

Suppose we are given the distribution $P_{\eta\xi_1,\xi_2\ldots\xi_n}$ and a subset $a = \{\xi_{i_1}, \xi_{i_2}, \cdots \xi_{i_k}\}$ of the set $\{\xi_1, \xi_2, \cdots \xi_n\}$ of all *symptom variables*. (Subset a is called *aperture* to stress it is a kind of filtering window through which we can see values of some symptom variables only during the decision making.) Then, the diagnostic problem can be formulated in the following way:
Diagnostic problem Find the diagnosis $d_a(s_{i_1}, s_{i_2} \cdots s_{i_k})$ that is the most probable (according to the $P_{\eta\xi_1,\xi_2\cdots\xi_n}$) on the set

$$\{\omega \in \Omega \,|\, \xi_{i_1}(\omega) = s_{i_1} \,\&\, \xi_{i_2}(\omega) = s_{i_2} \,\&\, \cdots \xi_{i_k}(\omega) = s_{i_k}\} \qquad (1)$$

for a given (i.e. observed) arbitrary combination $(s_{i_1}, s_{i_2} \cdots s_{i_k})$ of values of *symptom variables* from the set a.

From theoretical point of view, the optimal diagnosis $d_a(s_{i_1}, s_{i_2} \cdots s_{i_k})$ is given as

$$d_a(s_{i_1}, s_{i_2} \cdots s_{i_k}) = \operatorname*{argmax}_{d \in \mathrm{R}(\eta)} P_{\eta\,|\xi_{i_1}\xi_{i_2}\ldots\xi_{i_k}}(d|s_{i_1}, s_{i_2} \cdots s_{i_k}) \qquad (2)$$

Unfortunately, in the "real world", we are never given the theoretical distribution $P_{\eta\xi_1\xi_2\cdots\xi_n}$ in full and directly. To compensate for this, we expect to have some indirect information about $P_{\eta\xi_1\xi_2\cdots\xi_n}$ that will be called *knowledge base* and denoted by $\mathcal{K}$. It is done by postulating a set of conditions that we believe the theoretical $P_{\eta\xi_1\xi_2\cdots\xi_n}$ fulfills. Using the concept of *marginal problem*, see [1], *knowledge base* $\mathcal{K}$ is given as a set of "small-dimensional" distributions (i.e. number of variables in the distribution is small. E.g. not exceeding 10.), postulated as theoretical *marginal distributions* of the $P_{\eta\xi_1,\xi_2\ldots\xi_n}$. Instead of the unknown $P_{\eta\xi_1\xi_2\cdots\xi_n}$, we try to construct its suitable approximation $\hat{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ that could play its role in the diagnostic problem. Here, the small-dimensional distributions are either explicitly given or calculated from *statistical data file* T. This file T can be considered as realizations of the theoretical distribution $P_{\eta\xi_1\xi_2\cdots\xi_n}$. (Marginals that are derived from T can can be understood as marginals of the empirical distribution $P^T_{\eta\xi_1\xi_2\cdots\xi_n}$ given by T). There is an assumption that T is big enough so that marginals $P^T_{\eta\xi_{i_1}\xi_{i_2}\ldots\xi_{i_k}}$ of $P^T_{\eta\xi_1\xi_2\cdots\xi_n}$ can replace marginals $P_{\eta\xi_{i_1}\xi_{i_2}\ldots\xi_{i_k}}$ of the theoretical distribution $P_{\eta\xi_1\xi_2\cdots\xi_n}$. Instead of "small-dimensional distributions in $\mathcal{K}$", the one word term "oligodistributions" will be used in the sequel. This reflects the fact that they have usually a few of variables $s_{i_1}, s_{i_2} \cdots s_{i_k}$ and their respective sample spaces like $\mathrm{R}(\xi_l)$ consist of a few values only. (If the variables or sample spaces were not limited, though finite, there would be complexity problems with algorithms.) The second reason why the term *oligodistributions* is preferred to term *marginals* lies in the fact that small-dimensional distributions that are given by an expert as input need not be consistent and then, there does not exist any joint distribution whose marginals they might be. The third reason for the term *oligodistributions* is that we suppose it contains always diagnostic variable η though it is not mentioned explicitely in its carrier. (Carrier is the set of symptom variables in the

oligodistribution and it is denoted by underlining. E.g. if $o_i = P_{\eta\xi_{i_1}\xi_{i_2}\ldots\xi_{i_k}}$, then $\underline{o_i} = \{\xi_{i_1}\xi_{i_2}\ldots\xi_{i_k}\}$)
Decision making is based on information of two types. It is general knowledge about the problem area and specific evidence describing concrete patient(person). Knowledge is given by a set of oligodistributions (marginals), evidence is given by a vector of values that are taken by *symptom* variables from *aperture* $a \subseteq \Xi$ for the patient r_i.

There are two concepts that are indispensable for construction of the algorithm $\mathcal{SM}$ which are, however, *not* the topic of the paper. Namely, it is the notion of decision-making algorithm A_i (i.e. inference machine that solves the diagnostic problem) and, second, we need a testing scheme that evaluates the effectiveness of different A_i when they predict the diagnosis from symptoms for objects(persons) where the actual diagnosis is known. For the purpose of this paper, these two concepts can be modeled as two mappings, denoted as A_i and M, without going into details.

$$\begin{aligned} A_i : 2^{\Xi} \times Z(\Xi) \times \mathcal{R}_{\Xi} &\longrightarrow \mathcal{R}_{\eta} \\ (a, Z, r_s) &\longmapsto d_i \in \mathcal{R}_{\eta} \end{aligned}$$

where $Z = \{o_{i_1}, o_{i_2}, \cdot o_{i_k}\} \in Z(\Xi)$ is the knowledge base of the algorithm A_i, a is aperture and r_s is a concrete person from file T i.e. the realization r_s generated by $P_{\eta\xi_1\xi_2\cdots\xi_n}$ can be seen as a vector from $\mathcal{R}_{\eta\Xi}$

$r_s = (\eta(\omega_s), \xi_1(\omega_s), \xi_2(\omega_s) \cdot \xi_n(\omega_s))$, where $\omega_s \in \Omega$.
Examples of different algorithms A_i and the way they integrate the knowledge from Z can be found in [2].

To make the notation more compact, the symbolic mappings $\pi_s(.)$ can be used that, when applied to an object that is a vector (or set), returns the s-th component of the argument. E.g. $\pi_1(r_i) = \eta(\omega_i)$. Then, we may describe the testing scheme as a mapping M

$$\begin{aligned} M : \{A_i\}_i \times, 2^{\Xi} \times Z(\Xi) \times \{T_l\}_l &\longrightarrow \mathcal{I} \\ (A_i, a, Z, T) &\longmapsto |\{r_j \in T | A_i(a, Z, r_j) \neq \pi_1(r_j)\}| \end{aligned}$$

that for each inference algorithm A_i, equipped with a knowledge base Z, whose evidence information is restricted by an aperture a and that is applied to all persons r_i from the file T, returns the number of incorrect decisions (misclassifications) given by the condition $A_i(a, Z, r_j) \neq \pi_1(r_j)$. The mapping M (i.e. the testing scheme) can be easily coded and it is a part of program infrastructure (similarly as inference algorithms A_i or subroutine Comb) when $\mathcal{SM}$ is implemented. Using this formalism, we may proceed to the description of the $\mathcal{SM}$ algorithm.

2 Algorithm

2.1 Selecting Marginals

The **Selecting Marginals** algorithm ($\mathcal{SM}$ in the sequel) may be decomposed into two separate subtasks.
First, feasible marginals are selected.
Second, from the set of feasible marginals, the optimal knowledge base (i.e. set of the "best " marginals) is chosen.
In case of success, third step is added. It calculates the quality of decision making of inference algorithm A_i with the recommended knowledge base.

Inputs of $\mathcal{SM}$:

1. T is a statistical data file
2. a is the aperture (i.e.symptom variables unveiled for decision)
3. l is the requested size of the created knowledge base $\mathcal{K}$
4. c constant measures the "quality" of input marginals
5. A_i is the inference algorithm used for the third step

Outputs of $\mathcal{SM}$:

1. variable *error* reflects the final state of $\mathcal{SM}$ (i.e.$error = 0$ stands for success)
2. $\mathcal{K}$ is the optimal knowledge base
3. mc is number of misclassifications for $\mathcal{K}$, A_i and T.

The basic structure of algorithm $\mathcal{SM}$ can be the following one:

main **SelectingMarginals**(T, a, l, c, A_i, error, $\mathcal{K}$, mc)
  call **Feasiblemarginals**($a, c, T, SetFM, error$)
    if error = 1 then
      print "No marginals ! Reconsider c, T !"
      exit **SelectingMarginals**
    endif
  call **OptimalKnowledgeBase**($l, SetFM, \mathcal{K}$)
  call **FinalResults**($a, \mathcal{K}, A_i, T, mc$)
    print "Knowledge base $\mathcal{K}$ contains l marginals "
    print "With A_i, $\mathcal{K}$ yields ms misclassifications !"
end **SelectingMarginals**;

Subroutines **SelectingMarginals** and **OptimalKnowledgeBase** will be described later, **FinalResults** may look like this :

sub **FinalResults**($a, \mathcal{K}, A_i, T, mc$)
  $mc = M(Ai, a, \mathcal{K}, T)$
end **FinalResults**;

2.2 Feasible Marginals

The basic idea of **FeasibleMarginals** algorithm ($\mathcal{FM}$ in the sequel) is to find (and keep for further usage) only such marginals o_i that have an acceptable ratio data vs. space. This condition can be expressed in the form

$$|T|/|o_i| > c \tag{3}$$

Data represented by $|T|$ is the number of records in the file T and the symbol $|o_i|$ stands for the number of atoms of the sample space of all variables described by the marginal o_i. Let us remind that each oligodistribution o_i contains, by definition, always the diagnostic variable η, though it is not mentioned explicitly in its carrier $\underline{o_i}$. Hence,

$$|o_i| = |\eta| \cdot \prod_{\xi_j \in o_i} |\,\xi_j\,|$$

Let us stress that, this way, condition on feasibility of $o_j (\equiv P_{\eta \underline{o_j}} \equiv P_{\eta \xi_{j_1} \xi_{j_2} \cdots \xi_{j_k}})$ is easily evaluated without necessity to consider individual values $P_{\eta \xi_{j_1} \xi_{j_2} \cdots \xi_{j_k}}$
Namely, one could be stricter and require the file T to be representative enough to have some records (patients) for all atoms that are not "generic zeroes". ("Generic zeroes" in o_i are atomic events that can never occur as their probability is, by definition, zero. E.g. maternity for men.) In such a case, all $|o_i|$ values in o_i had to be calculated from the file T and, at maximum, $|o_i|$ tests of form

$$P_{\eta \xi_{j_1} \xi_{j_2} \cdots \xi_{j_k}})(d_i, s_{i_1}, s_{i_2}, \cdots s_{i_k}) > c_2$$

had to be performed. This would raise the computational complexity.
In the function **Cond** , described below, the former (i.e. simpler) condition (3) is used.
FeasibleMarginals algorithm constructs a set SetFM of all oligodistributions fulfilling the condition (3) and not dominated by other oligodistributions o_j with greater carrier $\underline{o_j}$ and also fulfilling the condition (3)

$$\mathcal{FM} = \{o_i \mid Cond(o_i) \quad \text{et} \bigvee_{o_j \,:\, |\underline{o_j}| > |\underline{o_i}|} \text{non} \quad Cond(o_j) \tag{4}$$

In certain sense, SetFM is a Pareto set with respect to the symptom variables in the carrier $\underline{o_i}$. The reason why we try for marginals with the greatest possible carrier is that they have greater discernment power than the smaller ones,.
All types of marginals that can be generated from $|a|$ variables constitute a lattice that has $2^{|a|}$ elements. A direct but cumbersome way to generate feasible marginals types would be to construct all carriers $\underline{o_i} \subseteq a$, throw out all carriers not fulfilling the condition (3) and then to throw out all carriers that can be dominated.
A more refined way was selected instead. Algorithms are described in a symbolic language that can be easily re-coded in a general purpose language (like VBasic or Fortran).
First, subroutine **prune** finds out whether diagnostic variable η has not too many possible values with respect to the cardinality of the file T. Then $\mathcal{SM}$ would finish with no marginals selected. Next, all ξ_i from a are tested for $|T| > |\eta| \cdot |\xi_i| \cdot c$. The failure of the test results in eliminating the respective

ξ_i from a. (In fact, acceptable ξ_i are put in the a_{new} set and original a is replaced by a_{new} on leaving subroutine **prune**. If $a_{new} = \emptyset$, then $\mathcal{SM}$ finishes without any selected marginal. The purpose of the function **average** is to set an "average" level in the lattice, that will be used as a starting point for further processing. Symbol *maxcard* is a variable that denotes maximal cardinality of symptom variables ξ_j that are in a.

$$maxcard = \underset{\xi_i \in a}{\text{argmax}} \ |\xi_i| \tag{5}$$

A histogram *hist* is constructed that describes for each $i \in < 2, maxcard >$ number of variables $\xi_j \in a$ that have cardinality i. The function **average** must return at least the value 2, as otherwise the program should have stopped before, due to the tests in the previous **prune** subroutine. Let us suppose we have a subroutine **Comb**(n,k) that generates all $\begin{pmatrix} n \\ k \end{pmatrix}$ combinations of numbers $\{1, 2, \cdots n\}$ and returns this set. (**Comb**(n,k) is not described in the sequel.) The *Set0* is filled with all $\begin{pmatrix} |a| \\ average \end{pmatrix}$ average-tuples created from variables contained in the set a.

To make the description of the algorithm more compact, some auxillary mappings m_a, coding ξ_i from different a to integers, are introduced:

$$\bigvee_{a \subseteq \Xi} m_a : a \longrightarrow < 1, |a| > \text{ such that } \bigvee_{\xi_i \xi_j \in a} \text{ if } i < j \text{ then } m_a(\xi_i) < m_a(\xi_j)$$

Then, m_a^{-1} denotes inversion of m_a and if the mappings $m_a()$, m_a^{-1} () are applied on a class of sets, it is supposed that they are applied on the elements of the sets, so that e.g. for $a = \{\xi_{15}, \xi_1, \xi_{21}\}$, and average = 2

$$\begin{array}{lcl} Comb\,(3,2) & = & \{\{1,2\},\{1,3\},\{2,3\}\} \\ m_a^{-1}(Comb\,(3,2)) & = & \{\{\xi_1,\xi_{15}\},\{\xi_1,\xi_{21}\},\{\xi_{15},\xi_{21}\}\}. \end{array}$$

This way, Set0 can be expressed like

$$Set0 = m_a^{-1}(Comb(|a|, average)) \tag{6}$$

All members $\underline{o_j}$ of Set0 (i.e. carriers of potential marginals) are tested via **Cond**. If the condition holds, all its immediate supersets are put in the SetU. If the testing condition does not hold, all the immediate subsets are put in the SetL. After the testing, the $\underline{o_j}$ is removed from Set0 so that $Set0 = \emptyset$ at the end of the cycle. If the $SetU$ is empty, $SetL$ will be tested in its turn. If the $SetU$ is not empty and contains some carriers of potential oligostributions, the whole procedure is repeated with the only exception. In case of not meeting the condition, nothing is put in the SetL. If SetL is empty, $\mathcal{FM}$ has finished and SetFM contains Pareto set of carriers of feasible marginals. If the set SetL contains some carriers, then $Set0 = SetL$; $SetL = \emptyset$ and the whole procedure is repeated with the exception that if the condition **Cond** is met, no filling of SetU follows, but respective $\underline{o_j}$ is added to SetFM. In case of not meeting conditions

in **Cond**, all subsets of the respective o_j are added to SetL and o_j is removed from Set0. If SetL is empty, the algorithm $\mathcal{FM}$ finishes and SetFM contains all greatest mutually non dominanat carriers of potential marginals.

main **FeasibleMarginals** $(a, c, T, SetFM, error)$
call **prune**(a,error); if error = 1 then go to end endif
k = average(a,error) if error = 1 then go to end endif
$Set0 = m_a^{-1}\,(Comb(|a|, k))$
$SetL = \emptyset$; $SetU = \emptyset$; $SetFM = \emptyset$; U = true; L = false
start:
for $\underline{o_j} \in Set0$
if Cond($\underline{o_j}$) then
if $\overline{\text{U}}$ then
if $a = \underline{o_j}$ then $SetFM = \{o_j\}$; error=0; go to end endif
call $increase(a, \underline{o_j}, SetU)$; go to middle
endif
if L then
$SetFM = SetFM \cup \{\underline{o_j}\}$; go to middle
endif
else
if L then
call $decrease(\underline{o_j}, SetL)$
endif
endif
middle:
next $\underline{o_j}$
if $\overline{SetU} \neq \emptyset$ then $Set0 = SetU; SetU = \emptyset$; go to start endif
if $SetL \neq \emptyset$ then
$Set0 = SetL; Set0 = \emptyset$;U = false; L = true; go to start
else
if $SetFM = \emptyset$ then error = 1 else error = 0 endif
endif
end:
end **FeasibleMarginals**;
function **average**$(a, error)$
for i $\in < 1, 10 >$ $\{hist(i) = 0\}$
for $\xi_i \in a$
$hist(|\xi_i|) = hist(|\xi_i|) + 1; maxcard = max(maxcard, |\xi_i|)$
next $\xi_i|$
for $i \in\ < 1, maxcard >$
if $i^{hist(i)} * |\eta| * c > |T|$ then
$average = i - 1$
$hist(|\xi_i|) = hist(|\xi_i|) + 1; maxcard = max(maxcard, |\xi_i|)$
endif
next i
if $average = 0$ then
print "Too many diagnoses !"; $Error = 1$; exit **average**
endif
end **average**;
sub **decrease**$(\underline{o}, SetL)$

```
  for j ∈ o
      if Cond(o\j) then
          SetL = SetL ∪ {o\j}
      endif
   next j
end decrease;
sub increase(a, o, SetU)
  for j ∈ a\o
      if Cond(o ∪ {j}) then
          SetU = SetU ∪ {o ∪ {j}}
      endif
   next j
end increase;
function Cond(o)
     space = |η|
     for ξ_j ∈ o   {space = space * |ξ_j|}  next ξ_j
     if |T|/space > c then Cond = true else Cond = false endif
end Cond;
sub prune(a, error)
      error = 0; a_new = ∅
      if |η| * c > |T| then
        print "Too many diagnoses !"; error = 1;  exit prune endif
    for ξ_j ∈ a
     if |ξ_j| * |η| * c ≤ |T| then a_new = a_new ∪ {ξ_j} endif
    next ξ_j
    if a_new = ∅ then  print "Too few data !"; error = 1 else a = a_new  endif
end prune;
```

2.3 Optimal Knowledge Base

Second part of the algorithm $\mathcal{SM}$ selects the "optimal" knowledge base $\mathcal{K}$ from the set SetFM that was prepared in the first part via subroutine **Feasible-Marginals**.

This optimization is in the form of the subroutine **OptimalKnowledgeBase** and it fills the set $\mathcal{K}$ with l oligodistributions for which the inference machines A_i decide with minimal number of misclassifications.

Basic idea of $\mathcal{OKB}$ consists in selecting oligodistributions with minimal number of misclassification. But, as the $|\mathcal{K}|$ is usually greater than 1, we calculate certain "correlation" $\rho(.,.)$ for all pairs of oligodistribution whose carriers are in SetFM.

$$\bigvee_{o_i,o_j:\underline{o_i},\underline{o_j}\in SetFM} \rho(o_i,o_j) = M(a, A_4, \{o_i,o_j\}, T) \tag{7}$$

Almost any inference algorithm A_i can be used for this calculation of M for a pair of oligodistributions, but we used A_4 from [3], as it is already in the program infrastructure.

Using this correlation ρ, we may construct the knowledge base $\mathcal{K}$ in an iterative way till the requested size l of $\mathcal{K}$ is achieved. Namely, we add a new oligodistribution $o_{|\mathcal{K}|+1}$ to the so far generated $\mathcal{K}$ if it is the most "orthogonal" to all

oligodistributions already present in $\mathcal{K}$. This is expressed by minimality of sum of correlations.

$$o_{|\mathcal{K}|+1} = \underset{o_s \notin \mathcal{K} : \underline{o_s} \in SetFM}{\operatorname{argmin}} \sum_{o_w \in \mathcal{K}} \rho(o_s, o_w) \tag{8}$$

```
 sub OptimalKnowledgeBase ((l, SetFM, K)
K = ∅
if  l = 1 then
    min = |T|; arg = 0
    for r = 1, |SetFM|
        if M (a, A_i, {o_r}, T) < min then
          min = M (a, A_i, {o_r}, T); arg = r
         endif
    next r
  K ={o_arg}
else
    SetP = Comb(|SetFM|, 2)
    for r = 1, l
    pair = π_r(SetP)
        if r = 1 then
          min = |T|; arg = 0
            for s = 1, ( |SetFM| 2 )
              m_s = M(a, A_4{π_1(pair), π_2(pair)}, T)
              if m_s < min then min = m_s; arg = s  elseif
          next s
          pair = π_arg(SetP)
          K = {π_1(pair), π_2(pair)}
        elseif r = 2 then
        else
          sum_min = |T|; arg = 0
          for u = 1, |SetFM|
           if o_u ∈ K then
           else
              sum = 0
              for t = 1, |K|
                  sum = sum + M(a, A_i, {o_u, π_t(K)}, T)
              next t
              if sum < sum_min then sum_min = sum; arg = u endif
           endif
          next u
          K = K ∪ {o_arg}
        endif
    next r
endif
end OptimalKnowledgeBase
```

3 Experimental results

$\mathcal{SM}$ algorithm was tested on the data from the field of rheumatology (prof.Rejholec, IV. Internal Clinic, I. Faculty of Medicine, Charles University, Czech republic, 1980). The *data file* T consists of 1089 patients and diagnosis variable η takes 4 different diagnoses. The file contains besides η, other 34 symptom variables ξ_i. To give better insight in this diagnostic problem, let us mention semantical meaning of some symptom variables. E.g. ξ_1 is sex, ξ_2 are age groups. ξ_3 stands for maternity maternity, ξ_{27} diabetes .... In this example, aperture a consists of eight symptom variables $a = \{\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, \xi_7, \xi_8\}$. Requested size l of final $\mathcal{K}$ was chosen to be 21. Subroutine FeasibleMarginals was skipped and we decided to generate all oligodistributions with four symptom variables that are in the aperture. Hence, SetFM was filled with $\begin{pmatrix} 8 \\ 4 \end{pmatrix}$ i.e. 70 oligodistributions as can be seen in the listing.

```
 1/ 8 over 4:   1   2   3   4
 2/ 8 over 4:   1   2   3   5
 3/ 8 over 4:   1   2   4   5
 ................
68/ 8 over 4:   3   6   7   8
69/ 8 over 4:   4   6   7   8
70/ 8 over 4:   5   6   7   8
```

There were $\begin{pmatrix} 70 \\ 2 \end{pmatrix}$ i.e. 2415 pairs of oligodistributions created from those 70 oligodistributions. The listing contains the pairs after sorting. E.g the second best pair has scored 204 misclassifications (out of 1089 cases), it consists from o_{33} and o_{38}. Oligoditribution o_{33}, when taken alone, yields 364 misclassifcations and o_{38}. achieves 637.

```
   1    196    14    38   368   637
   2    204    33    38   364   637
   3    211     4    33   605   364
   4    211     8    33   598   364
   5    211    14    21   368   599
   6    212    12    38   405   637
   7    212     4    14   605   368
   8    212    14    59   368   616
  ............................
2413    562    59    61   616   619
2414    575    57    59   620   616
2415    581    18    59   628   616
```

Finally, subroutine **OptimalKnowledgeBase** has selected the following 21 oligodistributions (out from the original 70): $o_{14}, o_{38}, o_{33}, o_{30}, \cdots o_{63}, o_9$. Second column is cumulative ρ correlation, fourth column is this sum of misclassifications divided by size of $\mathcal{K}$ in the l-th iterative step. The carrier $\underline{o_{14}}$ consists of $\xi_2, \xi_4, \xi_5, \xi_6$

```
1/    196     14    196 |    2   4   5   6
2/    196     38     98 |    1   3   4   8
```

3/	450	33	150	2	5	6	7
4/	698	30	174	2	4	6	7
5/	984	3	196	1	2	4	5
6/	1288	26	214	1	2	6	7
7/	1531	15	218	3	4	5	6
8/	1801	24	225	2	4	5	7
9/	2073	12	230	2	3	5	6
10/	2385	50	238	2	4	6	8
11/	2664	17	242	1	2	4	7
12/	2906	53	242	2	5	6	8
13/	3194	32	245	1	5	6	7
14/	3482	7	248	1	2	4	6
15/	3830	5	255	2	3	4	5
16/	4089	67	255	2	6	7	8
17/	4385	10	257	1	2	5	6
18/	4672	4	259	1	3	4	5
19/	5011	28	263	2	3	6	7
20/	5316	63	265	2	5	7	8
21/	5592	9	266	2	3	4	6

4 Conclusion

1. The idea not to apply *passively* just the marginals suggested by experts, but to look *actively* for the most appropriate knowledge base seems to be fruitful, as it is able to cut down misclassifications by up to 20 %. It is clear that requirement of statistical file T is nothing extraordinary as even with prescribed marginals one can hardly believe they could be obtained otherwise than from data.

2. The $\mathcal{SM}$ is challenging both from the point of view of program infrastructure as well as from time and space limits. These limits should be estimated for typical real situations. So far, it seems to be tolerable for $\mathcal{SM}$ to run several hours and precalculate the structure of database for different apertures a corresponding to batteries of tests.

3. More experiments should be organized, but it seems to be obvious that marginals constructed from variables ξ_i with larger ranges $\mathcal{R}(\xi_i)$ are preferred by $\mathcal{SM}$. Therefore, the role of constant c should be investigated in more detail. E.g. what are the acceptable ranges for c and what is the shape of functional dependence of misclassifications on c? Is it quasi linear?

4. It seems that successful inference algorithms A_i from [2] behave similarly with the same knowledge base. Then, improving its composition via $\mathcal{SM}$ is beneficial independently of A_i selected for routine prediction.

5. The number of marginals in the knowledge base may be a problem for some A_i. On the other hand, an increase in the parameter l decreases misclassifications. Is there an optimal limit for l ?

References

[1] Kellerer, H.G. (1964) Verteilungsfunktionen mit gegebenem Marginalverteilungen (in German), Z. Wahrsch. Verw. Gebiete, 3, 247-270.

[2] Kříž, O. (2007) Comparing algorithms based on marginal problem, Kybernetika, 43, 633-647.

[3] Kříž, O.: A new algorithm for decision making with probabilistic background, in: Transactions of the Eleventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, August 27-31, 1990, Vol. B, (Academia,Prague,1992) pp 135-143

Belief Functions on Formulas in Łukasiewicz Logic

Tomáš Kroupa*
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, 182 08 Prague
Czech Republic
kroupa@utia.cas.cz

Abstract

Belief functions are generalized to formulas in Łukasiewicz logic. It is shown that they generalize probabilities on formulas (so-called states) and that they are completely monotone mappings with respect to the lattice operations.

1 Introduction

Belief measures are certain non-additive real-valued set functions introduced by Dempster and Shafer [10, 12]. Roughly speaking, models based on belief measures are used in situations in which the precise probabilistic model consisting of one probability measure is not available due to the lack of information about the conditions or results of some random experiment. From the mathematical point of view, belief measures are completely monotone set functions in the sense of Choquet [11], who studied complete monotonicity of capacities in the systematic way.

The aim of this paper is to introduce belief functions in the framework of Łukasiewicz logic. This is accomplished by an extension procedure that assigns a functional to some belief measure via Choquet integral. In this general setting the key issue is to clarify the meaning of total monotonicity, which can be expressed on an arbitrary Abelian semigroup according to Choquet. The concept of belief function proposed in this paper includes many-valued analogues of probabilities on formulas, the so-called states. States were introduced by Mundici [9] in order to model the notion of "average truth-value" of formulas. It was proved in [5] and [6] that the mathematical properties of states indeed fits this idea, namely, every state is the Lebesgue integral of (an equivalence class of) a formula w.r.t. a Borel probability measure on possible worlds. Since this result is of an independent interest and motivates the forthcoming definition of a belief function, a new proof is given in Section 3.

*The work of the author was supported by the grant GA ČR 201/09/1891 and by the grant No.1M0572 of the Ministry of Education, Youth and Sports of the Czech Republic.

The paper is structured as follows. Section 2 contains necessary definitions and results concerning Łukasiewicz infinite-valued propositional logic and its associated Lindenbaum algebra L_k of (equivalence classes of) formulas over k propositional variables. Section 3 is devoted to states. In particular, it will be shown that the geometrical structure of formulas in L_k makes possible to derive the integral representation of states (Theorem 1). In Section 4 we investigate belief functions on formulas in L_k and show a number of generalizations of results known for classical belief measures on events (Theorem 3 and 4).

2 Preliminary Notions

The aim of this section is to provide a survey of Łukasiewicz infinite-valued propositional logic [1, Chapter 4] and its associated Lindenbaum algebra. Formulas $\varphi, \psi, \ldots$ are constructed from propositional variables $A_1, \ldots, A_k$ by applying the standard rules known in Boolean logic. The connectives are negation, disjunction and conjunction, which are denoted by $\neg$, $\oplus$ and $\odot$, respectively. This is already a complete set of connectives so that, for instance, the implication $\varphi \to \psi$ can be defined as $\neg\varphi \oplus \psi$. The set of all formulas in propositional variables $A_1, \ldots, A_k$ is denoted by $\mathrm{Form}(A_1, \ldots, A_k)$.

Semantics for connectives of Łukasiewicz logic is defined by operations in algebras called MV-algebras [1]. The algebra of truth degrees of Łukasiewicz logic is the *standard MV-algebra*, which is the unit interval $[0, 1]$ endowed with the operations $\neg, \oplus, \odot$ defined as follows:

$$\begin{aligned} \neg a &= 1 - a \\ a \oplus b &= \min\{a + b, 1\} \\ a \odot b &= \max\{a + b - 1, 0\} \end{aligned}$$

A *valuation* is a mapping $V : \mathrm{Form}(A_1, \ldots A_k) \to [0, 1]$ such that $V(\neg\varphi) = 1 - V(\varphi)$, $V(\varphi \oplus \psi) = V(\varphi) \oplus V(\psi)$ and $V(\varphi \odot \psi) = V(\varphi) \odot V(\psi)$. Formulas $\varphi, \psi \in \mathrm{Form}(A_1, \ldots, A_k)$ are called *equivalent* when $V(\varphi) = V(\psi)$, for every valuation V. The *equivalence class* of φ is denoted $[\varphi]$. The set of all such equivalence classes is an MV-algebra L_k with the operations $\neg[\varphi] = [\neg\varphi]$, $[\varphi] \oplus [\psi] = [\varphi \oplus \psi]$ and $[\varphi] \odot [\psi] = [\varphi \odot \psi]$, for every $\varphi, \psi \in \mathrm{Form}(A_1, \ldots, A_k)$.

Since every valuation V is uniquely determined by its restriction to the propositional variables $V \mapsto V(A_1, \ldots, A_k) \in [0, 1]^k$, every "possible world" V is matched with a unique point x_V from the k-dimensional unit cube $[0, 1]^k$ and vice versa. Let V_x be the valuation corresponding to $x \in [0, 1]^k$. Put $[\varphi](x) = V_x(\varphi)$, for every $x \in [0, 1]^k$. Hence the equivalence class $[\varphi]$ of every $\varphi \in \mathrm{Form}(A_1, \ldots A_k)$ can be viewed as a function $[0, 1]^k \to [0, 1]$ and L_k is the algebra of all such functions endowed with the pointwise operations $\neg, \oplus, \odot$.

McNaughton theorem ([2]). *$(L_k, \oplus, \odot, \neg)$ is precisely the algebra of all functions $[0, 1]^k \to [0, 1]$ that are continuous and piecewise linear, where each linear piece has integer coefficients.*

Let $f \vee g = \neg(\neg f \oplus g) \oplus g$, $f \wedge g = \neg(\neg f \vee \neg g)$. These operations are in fact the pointwise supremum and infimum of functions in L_k, respectively, and they make L_k into a distributive lattice.

A *filter* in L_k is a subset $\mathcal{F}$ of L_k such that (i) $1 \in \mathcal{F}$; (ii) if $f \in \mathcal{F}$ and $f \leq g$ with $g \in L_k$, then $g \in \mathcal{F}$; (iii) if $f, g \in \mathcal{F}$, then $f \odot g \in \mathcal{F}$. In this article we consider only filters $\mathcal{F}$ with $\mathcal{F} \neq L_k$. A *maximal filter* is a filter $\mathcal{F}$ such that no filter in L_k strictly contains $\mathcal{F}$.

Theory of Schauder hats and bases in L_k, which was developed for the purely geometrical proof of McNaughton theorem [1, Section 9.1], is briefly repeated in this paragraph. The basic familiarity with polyhedral geometry and topology is assumed, see [3, 4], for instance. A *polyhedral complex (in* $[0,1]^k$*)* is a finite set of polyhedra $\mathcal{R}$ such that: (i) each polyhedron of $\mathcal{R}$ is included in $[0,1]^k$, all its vertices have rational coordinates; (ii) if $P \in \mathcal{R}$ and Q is a face of P, then $Q \in \mathcal{R}$; (iii) if $P, Q \in \mathcal{R}$, then $P \cap Q$ is a face of both P and Q. The set $\bigcup_{P \in \mathcal{R}} P$ is called a *support* of $\mathcal{R}$. When all the polyhedra of a polyhedral complex $\mathcal{S}$ are simplices, then $\mathcal{S}$ is said to be a *simplicial complex*. Alternatively, a simplicial complex $\mathcal{S}$ with the support S is called a *triangulation* of S. The *denominator* $\mathrm{den}(q)$ of a point $q \in [0,1]^k$ with rational coordinates $(\frac{r_1}{s_1}, \ldots, \frac{r_k}{s_k})$, where $r_i \geq 0, s_i > 0$ are the uniquely determined relatively prime integers, is the least common multiple of $s_1, \ldots, s_k$. Passing to homogeneous coordinates in $\mathbb{R}^k$, put

$$\tilde{q} = \left(\tfrac{\mathrm{den}(q)}{s_1} r_1, \ldots, \tfrac{\mathrm{den}(q)}{s_k} r_k, \mathrm{den}(q)\right)$$

and note that $\tilde{q} \in \mathbb{Z}^{k+1}$. A k-simplex with vertices $v^0, \ldots, v^k$ is *unimodular* if $\{\tilde{v}^0, \ldots, \tilde{v}^k\}$ is a basis of the free Abelian group $\mathbb{Z}^{k+1}$. An n-simplex with $n < k$ is *unimodular* when it is a face of some unimodular k-simplex. We say that a triangulation Σ is *unimodular* if each simplex of Σ is unimodular. When $\mathcal{R}$ is a polyhedral complex, $\mathrm{V}(\mathcal{R})$ denotes the set of all the vertices of $\mathcal{R}$. Let Σ be a unimodular triangulation with a support $S \subseteq [0,1]^k$. For each $x \in \mathrm{V}(\Sigma)$, the *Schauder hat (at* x *over* Σ*)* is the uniquely determined continuous piecewise linear function $h_x : S \to [0,1]$ which attains the value $\frac{1}{\mathrm{den}(x)}$ at x, vanishes at each vertex from $\mathrm{V}(\Sigma) \setminus \{x\}$, and is a linear function on each simplex of Σ. The *basis* H_Σ (over Σ) is the set $\{h_x \mid x \in \mathrm{V}(\Sigma)\}$.

3 States

States on MV-algebras are many-valued analogues of probabilities on Boolean algebras. The disjointness of functions in L_k is captured by the relation $f \odot g = 0$, for $f, g \in L_k$. This condition also implies $f \oplus g = f + g$.

Definition 1. *A* state s *on* L_k *is a mapping* $s : L_k \to [0,1]$ *such that* $s(1) = 1$ *and* $s(f \oplus g) = s(f) + s(g)$*, for every* $f, g \in L_k$ *with* $f \odot g = 0$.

States on any (semisimple) MV-algebra were completely characterized in [5] and independently in [6] as integrals.

Theorem 1. *If* s *is a state on* L_k*, then there exists a uniquely determined Borel probability measure* μ *on* $[0,1]^k$ *such that* $s(f) = \int f \, \mathrm{d}\mu$*, for each* $f \in L_k$.

In the rest of this section we give an alternative, a purely geometrical proof of Theorem 1. By $\mathcal{M}^1$ we denote the convex set of all Borel probability measures on $[0,1]^k$, which is a compact metric space in w^*-topology. For every sequence (μ_n) in $\mathcal{M}^1$,

$$\mu_n \xrightarrow{w^*} \mu \text{ iff } \int f d\mu_n \longrightarrow \int f d\mu,$$

for every continuous function $f : [0,1]^k \to \mathbb{R}$. Let s be a state on L_k. In the sequel $\mathfrak{T}$ denotes the collection of all unimodular triangulations of $[0,1]^n$. Theorem 1 will be established in three steps.

Claim 1. *For every $\Sigma \in \mathfrak{T}$, the set of Borel probability measures*

$$\mathcal{M}_\Sigma = \{\mu \mid s(h_x) = \textstyle\int h_x d\mu, \text{ for each } h_x \in H_\Sigma\}$$

is nonempty and w^*-closed.

Proof. Let δ_x denotes the Dirac measure concentrated at a point $x \in [0,1]^n$. Put

$$\delta = \sum_{x \in \mathrm{V}(\Sigma)} \mathrm{den}(x) s(h_x) \delta_x,$$

and observe that $\mathrm{den}(x)s(h_x) = s(\mathrm{den}(x)h_x) \in [0,1]$ for each $x \in \mathrm{V}(\Sigma)$. The sum $\sum_{x\in \mathrm{V}(\Sigma)} \mathrm{den}(x)h_x$ is constantly equal to 1 since it is equal to 1 at every vertex of $\mathrm{V}(\Sigma)$ and every Schauder hat is linear over each simplex of Σ. This gives

$$\sum_{x \in \mathrm{V}(\Sigma)} \mathrm{den}(x) s(h_x) = \sum_{x \in \mathrm{V}(\Sigma)} s(\mathrm{den}(x)h_x) = s\left(\sum_{x \in \mathrm{V}(\Sigma)} \mathrm{den}(x) h_x \right) = s(1) = 1.$$

Hence δ is a convex combination of Borel probability measures and therefore itself a Borel probability measure. We will show that $\delta \in \mathcal{M}_\Sigma$. For each vertex $x' \in \mathrm{V}(\Sigma)$, we get

$$\begin{aligned} \int h_{x'} \,\mathrm{d}\delta &= \sum_{x \in \mathrm{V}(\Sigma)} \int \mathrm{den}(x) s(h_x) h_{x'} \,\mathrm{d}\delta_x = \sum_{x \in \mathrm{V}(\Sigma)} \mathrm{den}(x) s(h_x) h_{x'}(x) \\ &= \mathrm{den}(x') s(h_{x'}) h_{x'}(x') = \mathrm{den}(x') s(h_{x'}) \frac{1}{\mathrm{den}(x')} = s(h_{x'}). \end{aligned} \tag{1}$$

In order to show that $\mathcal{M}_\Sigma$ is w^*-closed, consider a sequence (μ_n) in $\mathcal{M}_\Sigma$ with $\mu_n \xrightarrow{w^*} \mu$, for some $\mu \in \mathcal{M}^1$. It follows that for each $h_x \in H_\Sigma$ we obtain $s(h_x) = \int h_x \,\mathrm{d}\mu_n \longrightarrow \int h_x \,\mathrm{d}\mu$. Hence $s(h_x) = \int h_x \,\mathrm{d}\mu$ and $\mu \in \mathcal{M}_\Sigma$. □

Claim 2. *The collection of subsets $(\mathcal{M}_\Sigma)_{\Sigma \in \mathfrak{T}}$ of $\mathcal{M}^1$ has the finite intersection property.*

Proof. Let $\mathfrak{T}' \subseteq \mathfrak{T}$ be nonempty and finite. We will show that $\bigcap_{\Sigma \in \mathfrak{T}'} \mathcal{M}_\Sigma \neq \emptyset$. First, we will show that every pair of bases $H_{\Sigma_1}, H_{\Sigma_2}$, where $\Sigma_1, \Sigma_2 \in \mathfrak{T}'$, has a joint refinement (that is, there exists a basis H such that both H_{Σ_1} and H_{Σ_2} are included in the MV-algebra generated by H). This is proved directly as follows. The triangulations Σ_1, Σ_2 have a joint subdivison (that is, there exists a triangulation of $[0,1]^k$ with the property that each of its simplices is included in some simplex of H_{Σ_1} or H_{Σ_2}) by taking all the intersections of simplices of H_{Σ_1} and H_{Σ_2}, and eventually triangulating the resulting polyhedral complex. This triangulation can be in turn subdivided to a unimodular triangulation $\Sigma^* \in \mathfrak{T}$ [7, Claim 2]. The joint refinement of the bases $H_{\Sigma_1}, H_{\Sigma_2}$ is then the basis H_{Σ^*}. The same argument straightforwardly applies to the finite set of bases $\{H_\Sigma \mid \Sigma \in \mathfrak{T}'\}$. Let $H_{\Sigma'}$ be the basis refining each basis H_Σ, $\Sigma \in \mathfrak{T}'$.

Precisely, if $\Sigma \in \mathfrak{X}'$, then for each $h_y \in H_\Sigma$ there exist uniquely determined nonnegative integers α_x, where $x \in \mathrm{V}(\Sigma')$, such that $h_y = \sum_{x\in\mathrm{V}(\Sigma')} \alpha_x h_x$. Put $\delta = \sum_{x\in\mathrm{V}(\Sigma')} \mathrm{den}(x) s(h_x)\delta_x$. It follows that

$$\int h_y \,\mathrm{d}\delta = \sum_{x\in\mathrm{V}(\Sigma')} \alpha_x \int h_x \,\mathrm{d}\delta = \sum_{x\in\mathrm{V}(\Sigma')} \alpha_x s(h_x),$$

where the last equality results from the calculation completely analogous to (1). Since $\sum_{x\in\mathrm{V}(\Sigma')} \alpha_x h_x \leq 1$, we obtain $\sum_{x\in\mathrm{V}(\Sigma')} \alpha_x s(h_x) = s\left(\sum_{x\in\mathrm{V}(\Sigma')} \alpha_x h_x\right) = s(h_y)$, and thus $\delta \in \bigcap_{\Sigma\in\mathfrak{X}'} \mathcal{M}_\Sigma$. □

Claim 3. *The intersection $\bigcap_{\Sigma\in\mathfrak{X}} \mathcal{M}_\Sigma$ contains a single element μ which satisfies $s(f) = \int f \,\mathrm{d}\mu$, for every $f \in L_k$.*

Proof. As $\mathcal{M}^1$ is w^*-compact and $(\mathcal{M}_\Sigma)_{\Sigma\in\mathfrak{X}}$ is a collection of w^*-closed subsets having the finite intersection property, the intersection $\bigcap_{\Sigma\in\mathfrak{X}} \mathcal{M}_\Sigma$ is nonempty. Every probability measure $\mu \in \bigcap_{\Sigma\in\mathfrak{X}} \mathcal{M}_\Sigma$ represents the state s. Indeed, given a McNaughton function $f \in L_k$, find $\Sigma^* \in \mathfrak{X}$ and the basis H_{Σ^*} such that $f = \sum_{x\in\mathrm{V}(\Sigma^*)} \alpha_x h_x$, for uniquely determined nonnegative integers α_x [1, Theorem 9.1.5]. It results that

$$\begin{aligned} s(f) &= s\left(\sum_{x\in\mathrm{V}(\Sigma^*)} \alpha_x h_x\right) = \sum_{x\in\mathrm{V}(\Sigma^*)} \alpha_x s(h_x) = \sum_{x\in\mathrm{V}(\Sigma^*)} \alpha_x \int h_x \,\mathrm{d}\mu \\ &= \int \sum_{x\in\mathrm{V}(\Sigma^*)} \alpha_x h_x \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu. \end{aligned}$$

It remains to show that $\bigcap_{\Sigma\in\mathfrak{X}} \mathcal{M}_\Sigma$ is a singleton. By the way of contradiction, assume that there are Borel probability measures $\mu, \nu \in \bigcap_{\Sigma\in\mathfrak{X}} \mathcal{M}_\Sigma$ such that $\mu \neq \nu$. The Borel subsets of $[0,1]^n$ are generated by the collection of all open (in the subspace Euclidean topology of $[0,1]^n$) (hyper)rectangles with rational vertices: indeed, every open subset of $[0,1]^n$ can be written as a countable union of such rectangles. As a consequence, [8, Theorem 3.3] yields that there exists an open rectangle $R \subseteq [0,1]^n$ with rational vertices and $\mu(R) \neq \nu(R)$.

Let $\mathcal{R}$ be the polyhedral complex consisting of all the faces of the closure $\overline{R}$ of R. Taking an arbitrary point $r \in R$ with rational coordinates, consider the stellar subdivision $\mathcal{R}'$ of $\mathcal{R}$ (see [4, p.15]). The polyhedral complex $\mathcal{R}'$ can be triangulated without introducing any new vertices [4, Proposition 2.9]. In turn, the resulting simplicial complex can be subdivided into a unimodular triangulation Σ of $\overline{R}$ with a possible introduction of new vertices (see [7, Claim 2], for example).

For each $v \in \mathrm{V}(\Sigma) \cap R$, let h_v be the Schauder hat at v over Σ, and define a function $f_v : [0,1]^n \to [0,1]$ by

$$f_v(x) = \begin{cases} h_v(x), & x \in \overline{R}, \\ 0, & \text{otherwise.} \end{cases}$$

When $f = \bigoplus_{v\in\mathrm{V}(\Sigma)\cap R} f_v$, then it follows directly from unimodularity of Σ and the definition of f_v that $f \in L_k$. In particular, note that $f(x)$ vanishes iff

$x \in [0,1]^n \setminus R$ and thus

$$\sup_{m\in\mathbb{N}} \bigoplus_{i=1}^{m} f = \chi_R, \tag{2}$$

where χ_R is the characteristic function of R. For every $m \in \mathbb{N}$, the function $\bigoplus_{i=1}^{m} f$ is an n-variable McNaughton function, and (2) together with Lebesgue's dominated convergence theorem leads to the equality

$$\mu(R) = \sup_{m\in\mathbb{N}} \int \bigoplus_{i=1}^{m} f \,\mathrm{d}\mu = \sup_{m\in\mathbb{N}} \int \bigoplus_{i=1}^{m} f d\nu = \nu(R),$$

which is the contradiction. □

The state space of L_k is a compact convex set. It can be completely described by its extreme boundary (Krein-Milman theorem), which is formed by the states $s_x : f \in L_k \mapsto f(x)$, for every $x \in [0,1]^k$. In addition, the set of all such states can be bijectively mapped onto the set of all maximal filters in L_k [9, Theorem 2.5] by the mapping $s_x \mapsto \mathcal{F}_x = \{f \in L_k \mid s_x(f) = 1\}$.

Theorem 2 ([9])**.** *The set $S(L_k)$ of all states on L_k is a compact convex subset of the product space $[0,1]^{L_k}$. The set of all extreme points of $S(L_k)$ equals $\{s_x \mid x \in [0,1]^k\}$, which is a closed subset of $S(L_k)$ whose elements are in one-to-one correspondence with maximal filters in L_k.*

4 Belief Functions

Belief measures introduced in Dempster-Shafer theory [10, 12] are particular completely (totally) monotone mappings in the sense of Choquet [11]. The complete monotonicity of a real function can be defined on an arbitrary Abelian semigroup. Let $(G, *)$ be an Abelian semigroup and β be a mapping $G \to \mathbb{R}$. Put $\Delta_a^* \beta(x) = \beta(x) - \beta(x * a)$, for every $x, a \in G$.

Definition 2. *A mapping $\beta : G \to \mathbb{R}$ is* completely monotone *if*

$$\Delta_{a_n}^* \cdots \Delta_{a_1}^* \beta(x) \geq 0 \tag{3}$$

for every $n \geq 1$ and every $x, a_1, \ldots, a_n \in G$.

A completely monotone, normalized and nonnegative real function on a family of sets equipped with $\cap$ is known as a belief measure (function) [12].

Definition 3 (Belief measure)**.** *Let $(G, *) = (\mathcal{A}, \cap)$, where $\mathcal{A}$ is a family of subsets of some nonempty set X closed w.r.t. finite intersections such that $\emptyset, X \in \mathcal{A}$. A completely monotone function $\beta : \mathcal{A} \to [0,1]$ with $\beta(X) = 1, \beta(\emptyset) = 0$ is called a* belief measure.

In case that $\mathcal{A}$ is even an algebra of sets, the condition (3) can be equivalently expressed for belief measures as follows:

$$\beta\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{\substack{I\subseteq\{1,\ldots n\}\\ I\neq\emptyset}} (-1)^{|I|+1} \beta\left(\bigcap_{i\in I} A_i\right),$$

for every $A_1, \ldots, A_n \in \mathcal{A}$. In this case the nonnegativity of the first two successive differences in (3) implies that β is a *monotone* and a *supermodular* set function, respectively, where the latter property means that

$$\beta(A_1 \cup A_2) + \beta(A_1 \cap A_2) \geq \beta(A_1) + \beta(A_2),$$

for every $A_1, A_2 \in \mathcal{A}$. In particular, note that every finitely additive probability measure on $\mathcal{A}$ is a belief measure due to the inclusion-exclusion principle.

A plain generalization of the classical notion of a belief measure from Definition 3 towards the MV-algebra of McNaughton functions L_k leads to considering the Abelian semigroup $(L_k, \odot)$ together with the differences defined by the operator $\Delta^{\odot}$. This approach, however, does not seem to give the "right" concept of a belief function on L_k since not every state is completely monotone w.r.t. $\Delta^{\odot}$. In fact it is possible to find a state s and McNaughton functions $f, g_1, g_2 \in L_k$ such that $\Delta^{\odot}_{g_2}\Delta^{\odot}_{g_1}s(f) < 0$. The lack of complete monotonicity is caused by the absence of distributivity of $\odot$ over $\oplus$ (and vice versa), which is in a clear contrast to the properties of the lattice operations $\vee$ and $\wedge$ on L_k. Yet the requirement of complete monotonicity for states is rather natural due to the linearity of every state (cf. Theorem 1) and consistency with the classical definition of belief measure, which covers finitely additive probabilities. An alternative definition of belief function on L_k is proposed in the next paragraph and it is shown how this concept relates to complete monotonicity w.r.t. the Abelian semigroup $(L_k, \wedge)$ together with the operator $\Delta^{\wedge}$.

In the sequel we consider belief measures on the family $\mathcal{C}$ of all closed subsets of $[0,1]^k$. In particular, a belief measure β on $\mathcal{C}$ is *outer regular (w.r.t. $\mathcal{C}$)* if $\beta(A) = \inf\{\beta(B) \mid B \in \mathcal{C} \text{ and } B \supseteq A\}$, for every $A \in \mathcal{C}$.

Definition 4 (Belief function)**.** *Let β be an outer regular belief measure on $\mathcal{C}$. A* belief function *$\hat{\beta}$ on L_k is given by*

$$\hat{\beta}(f) = \int_0^1 \beta(f^{-1}([t,1]))\,\mathrm{d}t, \quad f \in L_k. \tag{4}$$

Thus saying that "$\hat{\beta}$ is a belief function on L_k" is equivalent to the existence of an outer regular belief measure β on $\mathcal{C}$ so that $\hat{\beta}$ and β are related by the formula (4). The functional $f \mapsto \int_0^1 \beta(f^{-1}([t,1]))\,\mathrm{d}t$ is also called the *Choquet integral* of f w.r.t. β [13]. Every pre-image $f^{-1}([t,1])$ is a closed set in $[0,1]^k$ and $\beta(f^{-1}([t,1]))$ is thus well-defined. Since the function $t \mapsto \beta(f^{-1}([t,1]))$ is bounded and non-increasing on $[0,1]$ for a fixed β and $f \in L_k$, the integral on the right-hand side of (4) exists as the Riemann integral. Definition 4 bears a resemblance to the approach of Goubault-Larrecq in [14], where, on the other hand, belief measures are defined on the lattice of open subsets of a certain topological space. The preference of closed sets over opens is immaterial from the viewpoint of Choquet integration (4) and it will be justified only in the following. In a nutshell, closed subsets of $[0,1]^k$ correspond one-to-one to particular basic belief functions.

States are special belief functions according to Definition 4. Indeed, if an outer regular belief measure β satisfies

$$\beta(A \cup B) + \beta(A \cap B) = \beta(A) + \beta(B), \quad \text{for every } A, B \in \mathcal{C},$$

then β determines a unique regular Borel measure [11, V.26.6], and, consequently, the corresponding $\hat{\beta}$ is a state on L_k by Theorem 1 since the Choquet integral w.r.t. a measure is just the Lebesgue integral. Moreover, Choquet proved in [11, VII.52] that the integral in (4) preserves complete monotonicity of β when the lattice operations on the domain of $\hat{\beta}$ are employed. Precisely, the following statement holds true.

Theorem 3 ([11])**.** *Every belief function $\hat{\beta}$ is completely monotone w.r.t. the Abelian semigroup $(L_k, \wedge)$.*

Any belief function $\hat{\beta}$ thus satisfies the following properties that are jointly equivalent to its complete monotonicity:

(i) $\hat{\beta}$ is monotone,

(ii) for every $f_1, \ldots, f_n \in L_k$ with $n \geq 2$:

$$\hat{\beta}\left(\bigvee_{i=1}^{n} f_i\right) \geq \sum_{\substack{I \subseteq \{1,\ldots n\} \\ I \neq \emptyset}} (-1)^{|I|+1} \hat{\beta}\left(\bigwedge_{i \in I} f_i\right). \tag{5}$$

Further properties of belief functions on L_k are direct consequences of the well-known properties of Choquet integral (see [13]).

Proposition 1. *Let $\hat{\beta}$ be a belief function on L_k. Then for every $f, g \in L_k$:*

(i) $\hat{\beta}(0) = 0, \hat{\beta}(1) = 1$

(ii) *if $f \leq g$, then* $\hat{\beta}(f) \leq \hat{\beta}(g)$

(iii) *if $f \odot g = 0$, then* $\hat{\beta}(f \oplus g) \geq \hat{\beta}(f) + \hat{\beta}(g)$

(iv) $\hat{\beta}(f) + \hat{\beta}(\neg f) \leq 1$

(v) *$\hat{\beta}$ is a state iff β satisfies $\beta(A \cup B) + \beta(A \cap B) = \beta(A) + \beta(B)$, for every $A, B \in \mathcal{C}$*

(vi) *if $f \odot g = 0$ and there is no pair $x, y \in [0,1]^k$ with $f(x) < g(x), f(y) > g(y)$, then $\hat{\beta}(f \oplus g) = \hat{\beta}(f) + \hat{\beta}(g)$*

Basic examples of belief functions are minima of McNaughton functions over closed subsets of $[0,1]^k$.

Example 1. *Let $C \in \mathcal{C}$ be nonempty and*

$$b_C(f) = \min\{f(x) \mid x \in C\}, \quad f \in L_k.$$

Then b_C is a belief function since one can write $b_C = \widehat{\beta_C}$, where

$$\beta_C(A) = \begin{cases} 1, & C \subseteq A, \\ 0, & \text{otherwise}, \end{cases} \qquad A \in \mathcal{C},$$

is an outer regular belief measure on $\mathcal{C}$.

Theorem 4. *The set $B(L_k)$ of all belief functions on L_k is a compact convex subset of the product space $[0,1]^{L_k}$. The set of extreme points $\operatorname{ext} B(L_k)$ of $B(L_k)$ is closed, equals $\{b_C \mid C \in \mathcal{C}, C \neq \emptyset\}$, and it is in one-to-one correspondence with filters in L_k.*

Proof. It is known that the set $B(\mathcal{C})$ of all outer regular belief measures on $\mathcal{C}$ is a compact convex subset of the product space $[0,1]^{\mathcal{C}}$ and that the set of extreme points of $B(\mathcal{C})$ is closed and equals $\{\beta_C \mid C \in \mathcal{C}, C \neq \emptyset\}$ (see [11, VII.50]). The mapping $\beta \mapsto \hat{\beta}$ is an affine and a continuous mapping of $B(\mathcal{C})$ onto $B(L_k)$ since Choquet integration is continuous for a fixed integrand. Moreover, it is also injective, which can be deduced from another result of Choquet [11, p. 266]. The one-to-one correspondence between $\{b_C \mid C \in \mathcal{C}, C \neq \emptyset\}$ and the filters in L_k follows from [1, Section 3.4]: given b_C, put

$$\mathcal{F}_C = \{f \in L_k \mid f(x) = 1, \text{for every } x \in C\}. \tag{6}$$

Vice versa, if $\mathcal{F}$ is a filter in L_k, let

$$K_{\mathcal{F}} = \bigcap_{f \in \mathcal{F}} f^{-1}(1). \tag{7}$$

Compactness of $[0,1]^k$ and closedness of each $f^{-1}(1)$ gives that the closed set $K_{\mathcal{F}}$ is nonempty. The two mappings from (6)-(7) are mutually inverse since [1, Theorem 3.4.3(ii)] shows that $C = K_{\mathcal{F}_C}$, for every $C \in \mathcal{C}$. □

By Krein-Milman theorem, every belief function on L_k is thus in the closure of some convex hull formed by belief functions b_C. In particular, the usual integral reformulation of Krein-Milman theorem together with Theorem 4 admits to prove another integral representation of $\hat{\beta}$. The uniqueness part of the next theorem can be deduced from the similar result [11, VII.50.1] for $B(\mathcal{C})$ by using the fact that $\beta \mapsto \hat{\beta}$ is an affine homeomorphism.

Theorem 5. *If $\hat{\beta}$ is a belief function on L_k, then there exists a unique regular Borel probability measure μ on* $\operatorname{ext} B(L_k)$ *such that*

$$\beta(f) = \int_{\operatorname{ext} B(L_k)} b_C(f) \, \mathrm{d}\mu, \quad f \in L_k.$$

4.1 Remarks

Every belief function of the form b_C for some $C \in \mathcal{C}$ preserves finite minima:

$$b_C(f \wedge g) = b_C(f) \wedge b_C(g), \quad f, g \in L_k.$$

In general, every minimum-preserving function $b : L_k \to [0,1]$ with $b(0) = 1, b(1) = 1$ is a belief function. These functions are termed *necessity measures (functions)* and they were recently investigated on formulas of n-valued Łukasiewicz logic in [15].

Belief measures can be interpreted as certain lower probabilities. The corresponding upper probabilities are called *plausibility measures* in Dempster - Shafer theory. If $\mathcal{A}$ is an algebra of sets and $\beta : \mathcal{A} \to [0,1]$ is a belief measure, then the plausibility measure π is defined by $\pi(A) = 1 - \beta(A^C)$, for every $A \in \mathcal{A}$. Properties of plausibility measures are "dual" to those of belief measures so that the general theory can be developed for any of them. Plausibility functions on L_k are defined analogously: if b is a belief function on L_k, then the function $p(f) = 1 - b(\neg f), f \in L_k$, is called a *plausibility function.* Observe that it is the involutivity of Łukasiewicz negation that makes b and p dual to each other:

$$b(f) = b(\neg\neg f) = 1 - p(\neg f), \quad f \in L_k.$$

4.2 Open problems

The important open question is whether complete monotonicity of a real mapping on L_k is sufficient for its representation by the Choquet integral w.r.t. some belief measure on $\mathcal{C}$. Precisely, if $b : L_k \to [0,1]$ is such that $b(0) = 0, b(1) = 1$ and b is completely monotone w.r.t. $(L_k, \wedge)$, is it true that there exists an outer regular belief measure β on $\mathcal{C}$ satisfying $\hat{\beta} = b$?

Another question of interest is whether a belief function b on L_k is a "lower probability", that is, whether the equality

$$b(f) = \inf \{s(f) \mid s \text{ state with } s \geq b\}, \quad f \in L_k,$$

holds true or not.

References

[1] R. L. O. Cignoli, I. M. L. D'Ottaviano, and D. Mundici. *Algebraic foundations of many-valued reasoning*, volume 7 of *Trends in Logic—Studia Logica Library*. Kluwer Academic Publishers, Dordrecht, 2000.

[2] R. McNaughton. A theorem about infinite-valued sentential logic. *J. Symbolic Logic*, 16:1–13, 1951.

[3] G. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[4] C. P. Rourke and B. J. Sanderson. *Introduction to piecewise-linear topology*. Springer Study Edition. Springer-Verlag, Berlin, 1982.

[5] T. Kroupa. Representation and extension of states on MV-algebras. *Archive for Mathematical Logic*, 45(4):381–392, 2006.

[6] G. Panti. Invariant measures in free MV-algebras. *Communications in Algebra*, 36(8):2849–2861, 2008.

[7] D. Mundici. Bookmaking over infinite-valued events. *Internat. J. Approx. Reason.*, 43(3):223–240, 2006.

[8] P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[9] D. Mundici. Averaging the truth-value in Łukasiewicz logic. *Studia Logica*, 55(1):113–127, 1995.

[10] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press. Princeton, NJ, 1976.

[11] G. Choquet. Theory of capacities. *Ann. Inst. Fourier, Grenoble*, 5:131–295 (1955), 1953–1954.

[12] G. Shafer. Allocations of probability. *Ann. Probab.*, 7(5):827–839, 1979.

[13] D. Denneberg. *Non-additive measure and integral*, volume 27 of *Theory and Decision Library. Series B: Mathematical and Statistical Methods*. Kluwer Academic Publishers Group, Dordrecht, 1994.

[14] J. Goubault-Larrecq. Continuous capacities on continuous state spaces. In L. Arge et al., editor, *Proceedings of the 34th International Colloquium on Automata, Languages and Programming (ICALP'07)*, volume 4596 of *LNCS*, pages 764–776, Wrocław, Poland, July 2007. Springer.

[15] T. Flaminio, L. Godo, and E. Marchioni. On the logical formalization of possibilistic counterparts of states over n-valued Łukasiewicz events. *Journal of Logic and Computation*, doi:10.1093/logcom/exp012, 2009.

Fuzzy Previsions and Applications to Social Sciences

Antonio Maturo
Department of Social Sciences
University of Chieti-Pescara
amaturo@unich.it, antmat@libero.it

Aldo G.S. Ventre
Department of Culture of the Project and Benecon Research Centre
Second University of Napoli
aldoventre@yahoo.it

Abstract

An extension of the concept of fuzzy measure is introduced in an analogous way that coherent prevision is an extension of the finitely additive probability. To this purpose we deal with two new concepts: the fuzzy prevision of random numbers, as an extension to random numbers of the fuzzy measure, and the Archimedean g-operation as an extension to the set of real numbers of the Archimedean t-conorm. Moreover the concept of decomposable fuzzy prevision is introduced and an extension of the Weber [23] classification theorem is shown. Finally some applications to Social Sciences are sketched.

1 Introduction

The concept of *finitely additive probability* has been extended in [7] to the *coherent prevision*. While a finitely additive probability is a function defined in a set of events, a coherent prevision is defined in a set of random numbers. The events are particular random numbers with codomain contained in $\{0, 1\}$.

The advantage of such an extension consists in the possibility to replace the union of events with the sum of random numbers. Then the theory of the finitely additive probability is framed in the more general environment of the vector space of random numbers. In this way also a very useful geometrical interpretation is obtained, based on hyperplanes, convex sets and join spaces (see, e.g., [4], [7], [14], [15], [16], [17], [18], [19], [20], [21]).

A generalization of finitely additive probability is given by the *fuzzy measure*. The concept of fuzzy measure has been widely dealt with, among the others, by [1], [2], [17], [18], [22], [23]. As for the finitely additive probability, the codomain of a fuzzy measure is the real interval $[0, 1]$, but the *additivity* is replaced by the weaker condition of *monotonicity*.

Decomposable fuzzy measures with respect to t-conorms are considered in several book and papers, (see, e.g.,[1], [2], [10], [12], [17], [18], [22], [23]). These

will be recalled in the next Sec. and are a palatable compromise between the too much general concept of fuzzy measure and the very particular one of finitely additive probability. The additivity is replaced by the weaker property of additivity w. r. to a t-conorm.

Our present aim is to introduce the concept of *fuzzy prevision* as an extension of the concept of fuzzy measure, in an analogous way that coherent prevision is an extension of the finitely additive probability in [7]. Moreover we present the concept of *decomposable fuzzy prevision* that may be the happy medium between a coherent prevision and a fuzzy prevision.

To this purpose we introduce the new concept of *Archimedean g-operation* as an extension to the set of real numbers of the Archimedean t-conorm considered by many authors (see, e.g., [10], [17], [18], [22], [23]).

Moreover we introduce the concept of decomposable fuzzy prevision and we propose an extension of the Weber classification theorem [23] to fuzzy prevision. Properties of fuzzy previsions and Archimedean g-operations and their applications to Social Sciences are investigated.

2 Decomposable Fuzzy Measures

2.1 Fuzzy Measures

Let U be a set and $\mathcal{F}$ a family of subsets of U containing $\{\emptyset, U\}$.

Definition 1. Let us define *finitely monotonic fuzzy measure* on $\mathcal{F}$ every real function, $m : \mathcal{F} \to R$, such that:

FM1 $m(\emptyset) = 0$; $m(U) = 1$;

FM2 $\forall A, B \in \mathcal{F}, A \subseteq B \Rightarrow m(A) \leq m(B)$.

If $\mathcal{F}$ is an algebra and **FM2** is replaced by the stronger condition:

FM2P $\forall A, B \in \mathcal{F},\ A \cap B = \emptyset \Rightarrow m(A \cup B) = m(A) + m(B)$.

then m reduces to a *finitely additive probability* [7], [8].

A finitely monotonic fuzzy measure m is said to be a *fuzzy measure* [22] if:

FM3 for every monotonic sequence $\{A_n\}_{n \in N}$ of elements of $\mathcal{F}$,

$$\lim_n A_n = A \in \mathcal{F} \Rightarrow \lim_n m(A_n) = m(A).$$

If U is finite or, more generally, $\mathcal{F}$ is finite, then a finitely monotonic fuzzy measure on $\mathcal{F}$ is also a fuzzy measure [1], [2], [10], [18], [22], [23].

2.2 Archimedean t-conorms

Let us recall some definitions (see, e.g., [10], [12], [23]).

Definition 2. A binary operation $\perp$ on the real unit interval $[0, 1]$ is called a t-conorm if it is:

- increasing in each argument;
- associative;

- commutative;
- with 0 as neutral element.

A t-conorm $\perp$ is said to be *Archimedean* if it is continuous and $x \perp x > x$, $\forall x \in (0,1)$. An Archimedean t-conorm is called strict if it is strictly increasing in the open square $(0,1)^2$.

The following representation theorem holds:

Theorem 1. [12] A binary operation $\perp$ on $[0,1]$ is an Archimedean t-conorm if and only if there exists a strictly increasing and continuous function $g : [0,1] \to [0,+\infty]$, with $g(0) = 0$, such that

$$x \perp y = g^{(-1)}(g(x) + g(y)).$$

Function $g^{(-1)}$ denotes the pseudo-inverse of g, i.e.:

$$g^{(-1)}(x) = g^{-1}(\min(x, g(1))).$$

Moreover:

- $\perp$ is strict if and only if $g(1) = +\infty$;
- the function g, called an *additive generator* of $\perp$, is unique up to a positive constant factor.

2.3 Decomposable fuzzy measures

Let us recall some fundamental definitions and theorems [23].

Definition 3. Let U be a set and $\mathcal{F}$ an algebra of subsets of U. A fuzzy measure m on $\mathcal{F}$ is said to be *decomposable* w. r. to a t-conorm $\perp$, or $\perp$-decomposable, if:

$$A \cap B = \emptyset \Rightarrow m(A \cup B) = m(A) \perp m(B).$$

The following classification theorem holds:

Theorem 2. [23] (see also [17], [18]). If the operation $\perp$ in $[0,1]$ is a strict Archimedean t-conorm, then:

S $g \circ m : \mathcal{F} \to [0,+\infty]$ is an infinite additive measure, whenever m is a $\perp$-decomposable one.

If $\perp$ is a nonstrict Archimedean t-conorm, then $g \circ m$ is finite and one of the following cases occurs:

NSA $g \circ m : \mathcal{F} \to [0,+\infty]$ is a finite additive measure;

NSP $g \circ m$ is a finite set function which is only pseudo additive, i.e., if $\{A_n\}_{n \in \{1,2,\ldots,s\}}$ is a family of pairwise disjoint elements of $\mathcal{F}$, then:

$$(g \circ m)(\cup_{n=1}^{s} A_n) < g(1) \Rightarrow (g \circ m)(\cup_{n=1}^{s} A_n) = \sum_{n=1}^{s} (g \circ m)(A_n);$$

$$(g \circ m)(\cup_{n=1}^{s} A_n) = g(1) \Rightarrow (g \circ m)(\cup_{n=1}^{s} A_n) \leq \sum_{n=1}^{s} (g \circ m)(A_n).$$

3 Coherent Previsions

It is worth recalling some basic concepts related with random numbers and coherent prevision (see, e.g., [7], [3], [5], [6], [9], [14], [15], [16], [19]) as an extension to random numbers of the coherent subjective probability (see, e.g., [7], [8]), in order to have same useful points of reference to introduce the concept of fuzzy prevision.

Every *event* E can be identified with its characteristic function (see, e.g.,[7], [13]) $\chi_E : \{E, E^c\} \to \{0, 1\}$, such that $\chi_E(E) = 1, \chi_E(E^c) = 0$, where 0 and 1 are the *truth values* of E.

More in general, a *random number* is a function $\varphi : \Pi \to R$, where Π is a partition of the certain event. Then the events can be seen as the random numbers with range contained in $\{0, 1\}$, and the *fuzzy events* can be defined as random numbers with range contained in $[0, 1]$ (see, e.g., [13]).

The importance of the coherent prevision and its extensions to Decision Making is emphasized in many books (see, e. g., [7], [11]) and papers (see, e. g., [14], [15], [16], [19]).

Definition 4. [7] (see, also, [15], [16], [19]). Let S be a non empty set of random numbers. We define *prevision* on S a function $P : S \to R$ such that:

P1 $\forall a, b \in R, \forall X \in S,\ a \leq X \leq b \Rightarrow a \leq P(X) \leq b$ (*mean property*);

P2 $\forall X, Y \in S, X + Y \in S \Rightarrow P(X + Y) = P(X) + P(Y)$ (*additivity*).

We say that the prevision P is *coherent* if there exists an extension of P to the vector space $V(S)$ generated by S, i.e., the set of linear combinations of elements of S with coefficients on R.

Remark 1. A prevision P on S reduces to a *finitely additive probability* on S if every element X of S assumes only values belonging to the set $\{0, 1\}$. Moreover, if every element X of S assumes only values belonging to the real interval $[0, 1]$, P can be seen as a *probability of fuzzy events*.

Some important consequences of the previous definition are shown in the following theorems (see, e.g., [7], [14], [15], [16], [19]).

Theorem 4. If P is a coherent prevision on S, we have:

MO (*monotonicity*) $\forall X, Y \in S, X \leq Y \Rightarrow P(X) \leq P(Y)$;

EX (*linearity*) there is only an extension P^* of P to $V(S)$ and for every $n \in N, X_1, X_2, ..., X_n \in S$, $c_1, c_2, ..., c_n \in R$, we have:

$$P^*(c_1X_1 + c_2X_2 + ... + c_nX_n) = c_1P(X_1) + c_2P(X_2) + ... + c_nP(X_n). \quad (1)$$

Theorem 5. If P is a coherent prevision on S then for every $n \in N, X_1, X_2, ..., X_n \in S$, $c, c_1, c_2, ..., c_n \in R$ we have:

P3 $c_1X_1 + c_2X_2 + ... + c_nX_n \leq c \Rightarrow c_1P(X_1) + c_2P(X_2) + ... + c_nP(X_n) \leq c$;

P4 $c_1X_1 + c_2X_2 + ... + c_nX_n \geq c \Rightarrow c_1P(X_1) + c_2P(X_2) + ... + c_nP(X_n) \geq c$;

P5 $c_1X_1 + c_2X_2 + ... + c_nX_n = c \Rightarrow c_1P(X_1) + c_2P(X_2) + ... + c_nP(X_n) = c$.

Theorem 6. Let S be a non empty set of random numbers and let P be a function defined on S and with values on R. If, for every $n \in N, X_1, X_2, ..., X_n \in S$, $c, c_1, c_2, ..., c_n \in R$, **P3** or **P4** holds, then $P^* : V(S) \to R$ given by (1) is a prevision on $V(S)$ and so P is a coherent prevision on S.

4 Fuzzy Previsions

4.1 Fuzzy previsions as extension of fuzzy measures

The concept of prevision of random numbers can be extended from many different points of view. We consider two cases:

RNE *Random Numbers Extension.* Random numbers are replaced by random fuzzy numbers and fuzzy prevision is defined on a set of random fuzzy numbers with analogous properties as in definition 4. This way is pursued in some papers of ours (see, e.g., [14], [15], [16], [19]).

FME *Fuzzy Measure Extension.* The domain of a fuzzy measure is a set of events. We introduce a real function having as domain a set S of random numbers, such that its restriction to a set of events (i.e., random numbers with range contained in $\{0,1\}$) is a fuzzy measure. We pursue here this way.

Then we introduce the following definition.

Definition 5. Let S be a family of random numbers. We define *fuzzy prevision*, of type FME, on S, any function $P : S \to R$ such that:

FP1 $\forall a, b \in R, \forall X \in S,\ a \leq X \leq b \Rightarrow a \leq P(X) \leq b$; (*mean property*)

FP2 $\forall X, Y \in S, X \leq Y \Rightarrow P(X) \leq P(Y)$. (*monotonicity*)

4.2 An extension of the concept of Archimedean t-conorm

We introduce the following definition of additive generator on R, as a generalization of the concept of additive generator of an Archimedean t-conorm.

Definition 6. We define *additive generator* on R every function g defined in a closed interval $[a_g, b_g]$ of $[-\infty, +\infty]$, with codomain $[-\infty, +\infty]$, and such that:

AG1 the closed interval $[a_g, b_g]$ of $[-\infty, +\infty]$, called the *base interval*, contains $[0, 1]$;

AG2 $g(0) = 0$;

AG3 g is strictly increasing and continuous.

Definition 7. Let g be an additive generator on R with base interval $[a_g, b_g]$. We define *pseudoinverse* of g the function $g^{(-1)}$, defined in $[-\infty, +\infty]$, with codomain the base interval $[a_g, b_g]$ of g, and such that:

- $g^{(-1)}(y) = g^{-1}(y)$ if $y \in [g(a_g), g(b_g)]$;
- $g^{(-1)}(y) = a_g$ if $y \leq g(a_g)$;
- $g^{(-1)}(y) = b_g$ if $y \geq g(b_g)$.

Definition 8. Let g be an additive generator on R with base interval $[a_g, b_g]$. We define *Archimedean operation generated* by g, we call it the *g-operation*, the operation $\oplus$ defined as follows:

$$\forall x, y \in [a_g, b_g] : \{g(x), g(y)\} \neq \{-\infty, +\infty\},$$
$$x \oplus y = g^{(-1)}(g(x) + g(y)). \tag{2}$$

We say that the g-operation $\oplus$ is:

- *strict*, if $[g(a_g), g(b_g)] = [-\infty, +\infty]$;
- *nonstrict*, if $[g(a_g), g(b_g)]$ is a bounded interval of R;
- *semistrict*, otherwise.

From (2) the following theorem follows:
Theorem 7. The g-operation $\oplus$ given by (2) is:

- increasing in each argument;
- associative;
- commutative;
- with 0 as neutral element.

Moreover $\oplus$ is defined in $[a_g, b_g]^2$ if it is nonstrict or semistrict; while it is defined in $[a_g, b_g]^2 - \{(a_g, b_g), (b_g, a_g)\}$, if it is strict.
In particular, if $[a_g, b_g] = [0, 1]$, then $\oplus$ reduces to an Archimedean t-conorm.

4.3 Decomposable fuzzy previsions

Let us introduce a definition for *decomposable fuzzy prevision* as a generalization of decomposable fuzzy measure. To this aim the ambient *algebra of events* is replaced by the ambient *vector space*.

Let S be a vector space of random numbers, g an additive generator on R with base interval $[a_g, b_g]$, and $\oplus$ the correspondent g-operation. Moreover, let P be a fuzzy prevision on S with range $P(S)$ contained in $[a_g, b_g]$.

Definition 9. We say that P is a $\oplus$-*decomposable fuzzy prevision* on S if

$$\forall X, Y \in S : \{g(P(X)), g(P(Y))\} \neq \{-\infty, +\infty\}, \\ P(X + Y) = P(X) \oplus P(Y). \tag{3}$$

We shall prove two theorems, that provide for an extension of the Weber classification theorem [23] to the decomposable fuzzy previsions.

Theorem 8. (*Additivity theorem for strict g-operations*). If $\oplus$ is a strict g-operation and one of the following cases occur:

C1 $P(S) \subseteq [a_g, b_g)$;

C2 $P(S) \subseteq (a_g, b_g]$.

then $g \circ P$ is *additive*, that is:

$$\forall X, Y \in S, (g \circ P)(X + Y) = (g \circ P)(X) + (g \circ P). \tag{4}$$

Proof. From (2), (3) we have:

$$\forall X, Y \in S, g(P(X + Y)) = g(P(X) \oplus P(Y)) = g(g^{(-1)}(g(P(X)) + g(P(Y)))).$$

If **C1** or **C2** holds, then the sum $g(P(X)) + g(P(Y))$ is defined. Moreover, since $\oplus$ is a strict g-operation, $g^{(-1)}$ coincides with the inverse of g. Then

$$g(P(X+Y)) = g(P(X)) + g(P(Y)).$$

Theorem 9. (*Coherence theorem for nonstrict or semistrict g-operations*). If $\oplus$ is a nonstrict or semistrict g-operation, for every pair X, Y of random numbers belonging to S, we have:

A1 if $a_g < P(X+Y) < b_g$, then the *additivity* holds:

$$g(P(X+Y)) = g(P(X)) + g(P(Y)); \tag{5}$$

A2 if $P(X+Y) = a_g$, then we have the *superadditivity*:

$$g(P(X+Y)) \geq g(P(X)) + g(P(Y)); \tag{6}$$

A3 if $P(X+Y) = b_g$, then we have the *subadditivity*:

$$g(P(X+Y)) \leq g(P(X)) + g(P(Y)). \tag{7}$$

Proof. From (2), (3), $\forall X, Y \in S$ we have:

$$g(P(X+Y)) = g(P(X) \oplus P(Y)) = g(g^{(-1)}(g(P(X)) + g(P(Y))). \tag{8}$$

Let $z = g(P(X)) + g(P(Y))$. If z is not belonging to the open interval $(g(a_g), g(b_g))$ then $g(g^{(-1)}(z) \in \{(g(a_g), g(b_g)\}$. From (8) this implies that $g(P(X+Y)) \in \{(g(a_g), g(b_g)\}$, and so $P(X+Y) \in \{a_g, b_g\}$.

Then, in the case **A1**, $z \in (g(a_g), g(b_g))$, and then $g(g^{(-1)}(z) = z$, that is:

$$g(P(X+Y)) = g(P(X)) + g(P(Y)).$$

The cases **A2** and **A3** are an immediate consequence of the definition 7. Indeed, if $P(X+Y) = a_g$, then from (8) and definition 7 we have

$$g(P(X+Y)) = g(a_g),$$

$$g^{(-1)}(g(P(X)) + g(P(Y)) = a_g,$$

and then

$$g(P(X)) + g(P(Y)) \leq g(a_g) = g(P(X+Y)).$$

In an analogous way (7) is proved.

5 Applications of Fuzzy Previsions to Social Sciences

We consider the problem to build a Social and Cultural Center.

We assume there is a set $\mathcal{A} = \{A_1, A_2, ..., A_m\}$ of *alternative projects* and a set $\mathcal{O} = \{O_1, O_2, ..., O_n\}$ of *objectives* to be satisfied. It is reasonable to represent every objective O_j with a random number $X_j : \Pi \to R$, where Π is

the set of all the possible pairwise disjoint events and the range of X_j is the set of the possible *gains* or *utilities*.

A decision maker D associates to every pair (A_i, O_j) a real number P_{ij} that represents the prevision that D associates to X_j if the alternative A_i is realized.

It seems reasonable to assume the minimal requirement that every function:

$$P_i : O_j \in \mathcal{O} \to P_{ij}$$

is a *fuzzy prevision*.

In particular, if every P_i is decomposable w. r. to g-operation $\oplus$, we can assume that the global scores $S(A_i)$ of the alternatives A_i are obtained by the formula:

$$S(A_i) = P_{i1} \oplus P_{i2} \oplus ... \oplus P_{in}. \tag{9}$$

Of course, the choice of g-operation $\oplus$ depends on an in-depth study of the decision making problem and on the opinion of the decision maker D.

If there are more decision makers, each of these may have a different g-operation in order to aggregate the previsions of objectives.

References

[1] Banon G. (1981), Distinction between several subsets of fuzzy measures, *Int. J. Fuzzy Sets and Systems* 5, 291-305.

[2] Berres M. (1988), Lambda additive measure spaces, *Int. J. Fuzzy Sets and Systems* 27, 159-169.

[3] Berti P., Ragazzini E., Rigo P., (1994), *Coherent prevision of random elements*, CNR, Istituto per le applicazioni della matematica, Milano.

[4] Corsini P. and Leoreanu L. (2003), *Applications of the Hyperstructure Theory*, Kluver Academic Publishers, London.

[5] Crisma L., Gigante P., (2001), A notion of coherent conditional prevision for arbitrary random quantities, *Statistical Methods and Applications*, 10, 29-40.

[6] Crisma L., Gigante P., Millossovich P., (1997), A notion of coherent prevision for arbitrary random quantities, *Journal of the Italian Statistical Society*, vol 3, n 3, 233-243.

[7] de Finetti B. (1970), *Teoria delle probabilità*, Einaudi, Torino. Published in English as "Theory of Probability", (1974), J. Wiley, New York.

[8] Dubins L. E. (1975), Finitely additive conditional probabilities, conglomerability and disintegrations, *The Annals of Probability*, 3, 89-99.

[9] Holzer S., (1985), On coherence and conditional prevision, *Boll. UMI*, Ser. VI, Vol IV, 441-460.

[10] Klir G, Yuan B. (1995), *Fuzzy sets and fuzzy logic: Theory and Applications*, Prentice Hall, New Jersey.

[11] Lindley D. V., (1971), *Making decisions*, J. Wiley, New York.

[12] Ling C. H. (1965), Representation of associative functions, *Publ. Math. Debrecen* 12, 189-212.

[13] Maturo A. (2000), Fuzzy events and their probability assessments, *Journal of Discrete Mathematical Sciences & Cryptography*, Vol. 3, Nos 1-3, 83-94.

[14] Maturo A. (2003), On the conditional prevision of Bruno de Finetti and its applications. *6th Workshop on Uncertainty Processing*, Hejnice, September 24-27. September 24-27, 2003, 187-198.

[15] Maturo A. (2006), A geometrical approach to the coherent conditional probability and its fuzzy extensions, *Scientific Annals of University of A. S. V. M., "Ion Ionescu de la Brad"*, Iasi, XLIX, 2, 2006, 243-256.

[16] Maturo A. (2007), Algebraic hyperstructures and coherent conditional previsions. In: *Advances in Abstract Algebra*, Tofan, Gontineau, Tarnauceau Editors, pp. 1-18. A. Miller, IASI.

[17] Maturo A., Squillante M. and Ventre A. G. S., (2006) Consistency for assessments of uncertainty evaluations in non-additive setting, *in L. D'Ambra, P. Amenta, M. Squllante, A. G. S. Ventre, eds, Proceedings MTISD 06*, Franco Angeli, Milano.

[18] Maturo A., Squillante M., Ventre A. G. S. (2006), Consistency for non additive measures: analytical and algebraic methods. In: *B. Reusch, Computational Intelligence, Theory and Applications*, Springer-Verlag, Berlin, 29-40.

[19] Maturo A., Tofan I., Ventre A. (2004), Fuzzy Games and Coherent Fuzzy Previsions, in *Fuzzy Systems & A. I.*, 10, No 3, pp. 109-116.

[20] Maturo A., Ventre A. G. S. (2008), On Some Extensions of the De Finetti Coherent Prevision in a Fuzzy Ambit, *Journal of Basic Science 4*, No. 1(2008), 95-103

[21] Prenowitz W. and Jantosciak J. (1979), *Join geometries*, Springer Verlag VTM, New York.

[22] Sugeno M. (1974), *Theory of fuzzy integral and its applications*, Ph. D. Thesis, Tokyo.

[23] Weber S. (1984), Decomposable measures and integrals for Archimedean t-conorms, *J. Math. Anal. Appl.* 101 (1), 114-138.

Conditional Probability Spaces and Closures of Exponential Families

František Matúš
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
matus@utia.cas.cz

Abstract

A set of conditional probabilities is introduced by conditioning in the probability measures from an exponential family. A closure of the set is found, using previous results on the closure of another exponential family in the variational distance. The conditioning in the exponential family of all positive probabilities on a finite space is discussed and related to the permutahedra.

1 Introduction

A *conditional probability space* consists of a measurable space $(\Omega, \mathcal{A})$, nonempty set $\mathcal{B} \subseteq \mathcal{A}$ and family $\boldsymbol{P}$ of probability measures (pm's) $\boldsymbol{P}(\cdot|B)$, $B \in \mathcal{B}$, on $(\Omega, \mathcal{A})$ that satisfy $\boldsymbol{P}(B|B) = 1$ whenever $B \in \mathcal{B}$, and

$$\boldsymbol{P}(A|C) = \boldsymbol{P}(A|B) \cdot \boldsymbol{P}(B|C) \quad \text{whenever } A \in \mathcal{A},\ B, C \in \mathcal{B} \text{ and } A \subseteq B \subseteq C.$$

When viewed alternatively as a nonnegative function on $\mathcal{A} \times \mathcal{B}$, the family $\boldsymbol{P}$ is called the *conditional probability* (cp) on $(\Omega, \mathcal{A}, \mathcal{B})$ [12, 13, 8, 9, 3]. In this work, the set $\mathcal{B}$ is assumed to be finite.

Let μ be a finite nonzero measure on $(\Omega, \mathcal{A})$, $f\colon \Omega \to \mathbb{R}^d$ an $\mathcal{A}$-measurable function and $f\mu$ the image of μ under f, $f\mu(D) = \mu(f^{-1}(D))$, $D \subseteq \mathbb{R}^d$ Borel. The log-Laplace transform $\Lambda_{\mu,f}$ of $f\mu$,

$$\Lambda_{\mu,f}(\vartheta) = \ln \int_\Omega e^{\langle \vartheta, f\rangle}\, d\mu = \ln \int_{\mathbb{R}^d} e^{\langle \vartheta, x\rangle}\, f\mu(dx)\,, \qquad \vartheta \in \mathbb{R}^d\,,$$

is a convex function, finite on its nonempty domain $\mathit{dom}(\Lambda_{\mu,f})$ [14]. The full *exponential family* $\mathcal{E}_{\mu,f}$ determined by μ and f consists of the pm's $Q_{\mu,f,\vartheta}$ with the μ-density $e^{\langle \vartheta, f\rangle - \Lambda_{\mu,f}(\vartheta)}$ and $\vartheta \in \mathit{dom}(\Lambda_{\mu,f})$ [1, 2]. The family is endowed here with the topology of the variational distance $|P - Q|$ of pm's P and Q.

This work proposes to study sets of cp's that are analogous to the exponential families. Let μ be a measure on $(\Omega, \mathcal{A})$ that is positive and finite on $\mathcal{B}$, thus $0 < \mu(B) < +\infty$, $B \in \mathcal{B}$, and μ^B be the restriction of μ to B, $\mu^B(A) = \mu(A \cap B)$,

This work was supported by Grant Agency of Academy of Sciences of the Czech Republic, Grant IAA 100750603, and by Grant Agency of the Czech Republic, Grant 201/08/0539.

$A \in \mathcal{A}$. If ϑ belongs to $\mathit{dom}(\Lambda_{\mu^B,f})$ for every $B \in \mathcal{B}$ then the family $\boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta}$ of pm's defined by

$$\boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta}(A|B) = Q_{\mu^B,f,\vartheta}(A)\,, \qquad A \in \mathcal{A}\,, B \in \mathcal{B}\,,$$

is a cp on $(\Omega, \mathcal{A}, \mathcal{B})$ by Remark 2.1. The main object of interest here is the set

$$\mathfrak{E}^{\mathcal{B}}_{\mu,f} = \big\{\boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta} \colon \vartheta \in \textstyle\bigcap_{B\in\mathcal{B}} \mathit{dom}(\Lambda_{\mu^B,f})\big\}\,.$$

When $\mathcal{B} = \{\Omega\}$ this set is effectively the same as $\mathcal{E}_{\mu,f}$. It contains a cp $\boldsymbol{P}$ if and only if $\boldsymbol{P}(\cdot|\Omega) = Q_{\mu,f,\vartheta}$ for some $\vartheta \in \mathit{dom}(\Lambda_{\mu,f})$. Sets of cp's are endowed with the topology of the sum distance $\sum_{B\in\mathcal{B}} |\boldsymbol{P}(\cdot|B) - \boldsymbol{Q}(\cdot|B)|$ of cp's $\boldsymbol{P}$ and $\boldsymbol{Q}$.

Basic observations on the sets $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ are collected in Section 2. The main idea is to transform a cp $\boldsymbol{P}$ to the product of $\boldsymbol{P}(\cdot|B)$ over $B \in \mathcal{B}$, denoted by $\boldsymbol{\Pi P}$. The image $\boldsymbol{\Pi}\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ of $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ is recognized to be a full exponential family, see Lemma 2.3. This family is then reduced in two steps, see Lemma 2.4. A one-to-one canonical parametrization of the set $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ is described in Remark 2.7. Another parametrization follows from Lemma 2.9.

The closure of $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ is found in Theorem 3.3 applying the results of [6]. Under some assumptions on μ and f it is homeomorphic to a convex set, see Corollary 3.5.

Section 4 presents the special case of a finite Ω and the family $\mathcal{E}_{\mu,f}$ of all positive pm's on Ω. Relations to the algebraic approach of [11] are discussed. If $\mathcal{B}$ is the family of all nonempty subsets of Ω then the closure of $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ exhausts all cp's and can be parameterized by the points of a permutahedron of the dimension $|\Omega| - 1$, as found earlier in [10].

2 Basic observations

If a measure μ on $(\Omega, \mathcal{A})$ is positive and finite on $\mathcal{B}$ then the mapping

$$(A|B) \mapsto \tfrac{\mu(A\cap B)}{\mu(B)} = \tfrac{\mu^B(A)}{\mu^B(\Omega)}\,, \qquad A \in \mathcal{A}\,,\ B \in \mathcal{B}\,,$$

gives rise to a cp. For a necessary and sufficient condition on a cp to be generated from a measure as above see [4, (6.3), p. 351].

Remark 2.1. The assumption that μ is finite on $\mathcal{B}$ is equivalent to the finiteness of $\mu(\bigcup\mathcal{B})$ where $\bigcup\mathcal{B} = \bigcup_{B\in\mathcal{B}} B$, using that $\mathcal{B}$ is finite. If ν denotes the restriction of μ to $\bigcup\mathcal{B}$ then $\mathit{dom}(\Lambda_{\nu,f})$ is equal to the intersection of $\mathit{dom}(\Lambda_{\mu^B,f})$ over $B \in \mathcal{B}$. For ϑ in this domain

$$\tfrac{Q_{\nu,f,\vartheta}(A\cap B)}{Q_{\nu,f,\vartheta}(B)} = \textstyle\int_A e^{\langle\vartheta,f\rangle}\,d\mu^B \Big/ \int_\Omega e^{\langle\vartheta,f\rangle}\,d\mu^B = Q_{\mu^B,f,\vartheta}(A)\,, \qquad A \in \mathcal{A}\,,$$

thus $\boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta}$ is the cp on $(\Omega, \mathcal{A}, \mathcal{B})$ generated from $Q_{\nu,f,\vartheta}$. Hence, the set $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ can be constructed alternatively from $\mathcal{E}_{\nu,f}$ by conditioning to the sets $B \in \mathcal{B}$.

Lemma 2.2. *The mapping* $\boldsymbol{\Pi}$ *is a homeomorphism of the family of cp's into the family of product pm's on* $(\Omega^{\mathcal{B}}, \mathcal{A}^{\mathcal{B}})$.

Proof (sketch). The sum distance between cp's $\boldsymbol{P}$, $\boldsymbol{Q}$ majorizes the variational distance between $\boldsymbol{\Pi P}$ and $\boldsymbol{\Pi Q}$. The variational distance between two products of pm's majorizes the variational distance between any two marginal pm's. □

For a finite measure μ on $(\Omega, \mathcal{A})$, let $\mu_{\mathcal{B}}$ be product of the restrictions μ^B over $B \in \mathcal{B}$. For a function $f\colon \Omega \to \mathbb{R}^d$ let $f_{\mathcal{B}}$ map an element $\omega_{\mathcal{B}} = (\omega_B)_{B\in\mathcal{B}}$ of $\Omega^{\mathcal{B}}$ to $(f(\omega_B))_{B\in\mathcal{B}}$, an element of $(\mathbb{R}^d)^{\mathcal{B}}$. The function f is always assumed to be $\mathcal{A}$-measurable. Let Σ map $(x_B)_{B\in\mathcal{B}} \in (\mathbb{R}^d)^{\mathcal{B}}$ to $\sum_{B\in\mathcal{B}} x_B \in \mathbb{R}^d$.

Lemma 2.3. *If μ is positive and finite on $\mathcal{B}$ then*
(i) $\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}} = \sum_{B\in\mathcal{B}} \Lambda_{\mu^B,f}$
(ii) $\mathbf{\Pi}\boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta} = Q_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}},\vartheta}$ *for* $\vartheta \in \mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}})$
(iii) $\mathbf{\Pi}$ *restricts to a homeomorphism between* $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ *and* $\mathcal{E}_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}$.

Proof. For $\vartheta \in \mathbb{R}^d$

$$\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}(\vartheta) = \ln \int_{\Omega^{\mathcal{B}}} e^{\langle \vartheta, \Sigma f_{\mathcal{B}}\rangle}\, d\mu_{\mathcal{B}} = \ln \int_{\Omega^{\mathcal{B}}} \prod_{B\in\mathcal{B}} e^{\langle \vartheta, f(\omega_B)\rangle}\, \mu_{\mathcal{B}}(d\omega_{\mathcal{B}})$$

using $\langle \vartheta, \Sigma f_{\mathcal{B}}(\omega_{\mathcal{B}})\rangle = \sum_{B\in\mathcal{B}} \langle \vartheta, f(\omega_B)\rangle$. Hence,

$$\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}(\vartheta) = \ln \prod_{B\in\mathcal{B}} \int_{\Omega} e^{\langle \vartheta, f(\omega)\rangle}\, \mu^B(d\omega) = \sum_{B\in\mathcal{B}} \Lambda_{\mu^B,f}(\vartheta)$$

which proves *(i)*. It follows that $\mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}})$ is the intersection of $\mathsf{dom}(\Lambda_{\mu^B,f})$ over $B \in \mathcal{B}$. For ϑ in the domain the product pm $\mathbf{\Pi}\boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta}$ is absolutely continuous w.r.t. $\mu_{\mathcal{B}}$ and by *(i)* has the density

$$\begin{aligned}\prod_{B\in\mathcal{B}} d\boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta}(\cdot|B)\Big/ d\mu^B\,(\omega_{\mathcal{B}}) &= \prod_{B\in\mathcal{B}} \exp\left[\langle \vartheta, f(\omega_B)\rangle - \Lambda_{\mu^B,f}(\vartheta)\right] \\ &= \exp[\langle \vartheta, \Sigma f_{\mathcal{B}}(\omega_{\mathcal{B}})\rangle - \Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}(\vartheta)] = dQ_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}},\vartheta}\big/ d\mu_{\mathcal{B}}\,(\omega_{\mathcal{B}})\,,\end{aligned}$$

thus *(ii)* holds. Then *(iii)* follows by Lemma 2.2. □

Lemma 2.4. *If μ is positive and finite on $\mathcal{B}$ then*
(i) $\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}} = \Lambda_{f_{\mathcal{B}}\mu_{\mathcal{B}},\Sigma} = \Lambda_{\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}},\mathsf{id}}$
where id *denotes the identity mapping on $\mathbb{R}^d$, and for $\vartheta \in \mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}})$*
(ii) $f_{\mathcal{B}}Q_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}},\vartheta} = Q_{f_{\mathcal{B}}\mu_{\mathcal{B}},\Sigma,\vartheta}$
(iii) $\Sigma f_{\mathcal{B}}Q_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}},\vartheta} = \Sigma Q_{f_{\mathcal{B}}\mu_{\mathcal{B}},\Sigma,\vartheta} = Q_{\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}},\mathsf{id},\vartheta}$.

A proof is standard and omitted.

The *convex core* $\mathsf{cc}(\nu)$ of a finite Borel measure ν on $\mathbb{R}^d$ is intersection of the convex Borel sets $D \subseteq \mathbb{R}^d$ with $\nu(\mathbb{R}^d \setminus D) = 0$ [5]. Let $\mathsf{ri}(\nu)$ denote the relative interior of $\mathsf{cc}(\nu)$.

Lemma 2.5. *If μ is finite on $\mathcal{B}$ then*
(i) $\mathsf{cc}(f_{\mathcal{B}}\mu_{\mathcal{B}}) = \prod_{B\in\mathcal{B}} \mathsf{cc}(f\mu^B)$
(ii) $\mathsf{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}}) = \Sigma\mathsf{cc}(f_{\mathcal{B}}\mu_{\mathcal{B}}) = \sum_{B\in\mathcal{B}} \mathsf{cc}(f\mu^B)$.

Proof. Since $f_{\mathcal{B}}\mu_{\mathcal{B}}$ is the product of the measures $f\mu^B$ over $B \in \mathcal{B}$ the first equality follows from [5, Lemma 7]. The Σ-image of a product measure is the convolution of marginals. Hence, the second assertion is a consequence of [5, Corollary 8]. □

Corollary 2.6. *Lemma 2.5 remains valid when* cc *is replaced by* ri.

For a convex set $D \subseteq \mathbb{R}^d$ let $\mathsf{lin}(D)$ denote the linear space generated by the differences $x - y$ with $x, y \in D$ and π_D the orthogonal projection onto $\mathsf{lin}(D)$. In the case $D = \mathsf{cc}(\nu)$ the abbreviations $\mathsf{lin}(\nu)$ and π_ν are used.

Remark 2.7. If a measure μ on $(\Omega, \mathcal{A})$ is nonzero and finite, and $f\colon \Omega \to \mathbb{R}^d$ then $\vartheta \in \mathsf{dom}(\Lambda_{\mu,f})$ and $\pi_{f\mu}(\vartheta - \theta) = 0$ imply $\theta \in \mathsf{dom}(\Lambda_{\mu,f})$. The exponential family $\mathcal{E}_{\mu,f}$ is bijectively parameterized by $\pi_{f\mu}(\mathsf{dom}(\Lambda_{\mu,f}))$. If follows on account of Lemma 2.3 that $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ is bijectively parameterized by $\pi_{\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}}}(\mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}))$. Here, the projection is onto $\mathsf{lin}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$ which is the sum of $\mathsf{lin}(f\mu^B)$ over $B \in \mathcal{B}$, by Lemma 2.5 *(ii)*.

Remark 2.8. The log-Laplace transform $\Lambda_{\mu,f}$ is differentiable at any ϑ from the interior of its domain and $\nabla\Lambda_{\mu,f}(\vartheta) = \int_\Omega f\, dQ_{\mu,f,\vartheta}$ [1, 2]. If the domain is open then $\nabla\Lambda_{\mu,f}$ gives rise to a diffeomorphism between the relatively open sets $\pi_{f\mu}(\mathsf{dom}(\Lambda_{\mu,f}))$ and $\mathsf{ri}(f\mu)$. Thus, the mapping $P \mapsto \int_\Omega f\, dP$ is defined for every $P \in \mathcal{E}_{\mu,f}$, and it is a homeomorphism between $\mathcal{E}_{\mu,f}$ and $\mathsf{ri}(f\mu)$, see also [6, Corollary 1].

Let $\mathbf{M}_f$ denote the composition of two mappings

$$\boldsymbol{P} \mapsto \boldsymbol{\Pi P} \mapsto \int_{\Omega^{\mathcal{B}}} \Sigma f_{\mathcal{B}}\, d\,\boldsymbol{\Pi P}$$

defined at any cp $\boldsymbol{P}$ such that the integral exists. Rewriting the integral to

$$\int_{\Omega^{\mathcal{B}}} \sum_{B\in\mathcal{B}} f(\omega_B) \cdot \prod_{B\in\mathcal{B}} \boldsymbol{P}(d\omega_B|B)$$

the existence is equivalent to $\boldsymbol{P}(\cdot|B)$-integrability of f for $B \in \mathcal{B}$, in which case

$$\mathbf{M}_f\boldsymbol{P} = \sum_{B\in\mathcal{B}} \int_\Omega f(\omega)\, \boldsymbol{P}(d\omega|B)\,.$$

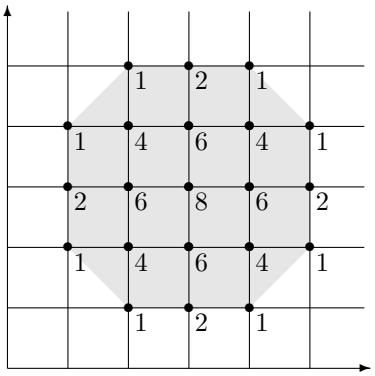
Lemma 2.9. *If* $\mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}})$ *is open then* $\mathbf{M}_f$ *restricts to a homeomorphism between* $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ *and* $\mathsf{ri}(\Sigma f_{\mathcal{B}}\, \mu_{\mathcal{B}}) = \sum_{B\in\mathcal{B}} \mathsf{ri}(f\mu^B)$.

Proof. The restriction is a composition of two homeomorphisms. The first one comes from Lemma 2.3 *(iii)*, between $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ and $\mathcal{E}_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}$. The second one makes homeomorphic $\mathcal{E}_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}$ and $\mathsf{ri}(\Sigma f_{\mathcal{B}}\, \mu_{\mathcal{B}})$, by Remark 2.8. It remains to refer to Corollary 2.6. □

Example 2.10. Let $\Omega = \{0,1\}^2$, $\mathcal{A}$ be the algebra of all subsets of Ω and $\mathcal{B} = \binom{\Omega}{2}$ consist of all two-element subsets of Ω. Let μ be the counting measure on Ω and f the embedding of Ω to $\mathbb{R}^2$. The family $\mathcal{E}_{\mu,f}$ consists of all positive product pm's on Ω and the set $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ of the cp's that are generated from these pm's, see Remark 2.1. Denoting by δ_x the Borel pm on $\mathbb{R}^2$ that is supported by $x \in \mathbb{R}^2$, the $f_{\mathcal{B}}$-image of $\mu_{\mathcal{B}}$ is the product

$$[\delta_{(0,0)}{+}\delta_{(1,0)}]\times[\delta_{(0,0)}{+}\delta_{(0,1)}]\times[\delta_{(0,0)}{+}\delta_{(1,1)}]\times[\delta_{(1,0)}{+}\delta_{(0,1)}]\times[\delta_{(1,0)}{+}\delta_{(1,1)}]\times[\delta_{(0,1)}{+}\delta_{(1,1)}]\,.$$

Then, $\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}}$ is the convolution of the six measures. It is equal to the linear combination of δ_x's where x runs over the points of the configuration below and the coefficients in the combination correspond to the labels of the points.



The shaded hexagon is the convex core of $\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}}$. Lemma 2.5 expresses the hexagon as the sum of the edges and diagonals of the unit square. Further,

$$\begin{aligned}\mathbf{M}_f\boldsymbol{P} &= \textstyle\sum_{B\in\mathcal{B}}\sum_{\omega\in\Omega} f(\omega)\,\boldsymbol{P}(\omega|B)\\ &= (1,0)[\boldsymbol{P}(10|00,10)+\boldsymbol{P}(10|10,01)+\boldsymbol{P}(10|10,11)]\\ &\quad +(0,1)[\boldsymbol{P}(01|00,01)+\boldsymbol{P}(01|10,01)+\boldsymbol{P}(01|01,11)]\\ &\quad +(1,1)[\boldsymbol{P}(11|00,11)+\boldsymbol{P}(11|10,11)+\boldsymbol{P}(11|01,11)]\end{aligned}$$

where e.g. $\boldsymbol{P}(10|00,10)$ is an abbreviation for $\boldsymbol{P}\big(\{(1,0)\}|\{(0,0),(1,0)\}\big)$. By Lemma 2.9, the mapping $\mathbf{M}_f$ restricts to a homeomorphism between $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ and the interior of the hexagon.

3 Closures of the families $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$

Given a convex set D in a Euclidean space, its nonempty convex subset F is a *face* if each segment contained in D with an interior point in F is contained in F.

Lemma 3.1. *If μ is finite on $\mathcal{B}$ and F is a face of $\mathsf{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$ then*
(i) $F_\Sigma = \Sigma^{-1}(F)\cap\mathsf{cc}(f_{\mathcal{B}}\mu_{\mathcal{B}})$ is a face of $\mathsf{cc}(f_{\mathcal{B}}\mu_{\mathcal{B}})$
(ii) $F_\Sigma = \prod_{B\in\mathcal{B}} F_{\Sigma,B}$ where $F_{\Sigma,B}$ is a unique face of $\mathsf{cc}(f\mu^B)$
(iii) $\Sigma F_\Sigma = F$.

Proof. The assertions follow from Lemma 2.5 and basic convex geometry. □

Let $\mu_{\mathcal{B},F} = \prod_{B\in\mathcal{B}}\mu^{B,F}$ where $\mu^{B,F}$ is the restriction of μ to $B\cap f^{-1}(\mathsf{cl}(F_{\Sigma,B}))$.

Lemma 3.2. *If μ is finite on $\mathcal{B}$ and F is a face of $\mathsf{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$ then $\mu_{\mathcal{B},F}$ is nonzero and finite and $\Sigma f_{\mathcal{B}}\mu_{\mathcal{B},F} = (\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})^{\mathsf{cl}(F)}$.*

Proof (sketch). Since F is a face, thus a nonempty set, every $F_{\Sigma,B}$ is a face of $\mathsf{cc}(f\mu^B)$ by Lemma 3.1. Therefore $f\mu^B(\mathsf{cl}(F_{\Sigma,B})) = \mu^{B,F}(\Omega)$ is positive by [5, Corolary 3]. Thus, $\mu_{\mathcal{B},F}$ is nonzero. Since μ is finite on $\mathcal{B}$ every $f\mu^B$ is finite, and the finiteness of $\mu_{\mathcal{B},F}$ follows.

The equality is a consequence of $f_{\mathcal{B}}\mu_{\mathcal{B},F} = (f_{\mathcal{B}}\mu_{\mathcal{B}})^{\Sigma^{-1}(\mathsf{cl}(F))}$. Since $f_{\mathcal{B}}\mu_{\mathcal{B},F}$ is the restriction of $f_{\mathcal{B}}\mu_{\mathcal{B}}$ to $\prod_{B\in\mathcal{B}}\mathsf{cl}(F_{\Sigma,B}) = \mathsf{cl}(F_\Sigma)$ the aim is to prove that $f_{\mathcal{B}}\mu_{\mathcal{B}}(\Sigma^{-1}(\mathsf{cl}(F))\setminus\mathsf{cl}(F_\Sigma)) = 0$, using that the two sets are in inclusion.

If $F_\Sigma = \mathsf{cc}(f_{\mathcal{B}}\mu_{\mathcal{B}})$ then $\mathsf{cl}(F_\Sigma)$ has the complement of $f_{\mathcal{B}}\mu_{\mathcal{B}}$-measure zero by [5, Lemma 1]. Otherwise, F is not equal to $D = \mathsf{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$. Assume first

that F is exposed, thus a nontrivial supporting hyperplane H to D exists such that $F = H \cap D$. Then, $\Sigma^{-1}(H)$ is a supporting hyperplane of $\mathsf{cc}(f_{\mathcal{B}}\mu_{\mathcal{B}})$ and $\Sigma^{-1}(H) \cap \mathsf{cc}(f_{\mathcal{B}}\mu_{\mathcal{B}}) = F_{\Sigma}$. By [6, Lemma 1], $f_{\mathcal{B}}\mu_{\mathcal{B}}(\Sigma^{-1}(H) \setminus \mathsf{cl}(F_{\Sigma})) = 0$ and the equality holds. If F is not exposed then it can be approached by a chain of exposed faces and the equation obtains from the corresponding equations in the chain. Details are omitted. □

Where D and Ξ are nonempty convex subsets in a Euclidean space, the concept of Ξ-*accessible* face of D was introduced in [6, Subsection 2.5]. The definition is rather technical and not repeated here, using later only the simple facts that D is always a Ξ-accessible face of D and every face is $\mathbb{R}^d$-accessible.

For a face F of $\mathsf{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$ the family $\boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta}$ of pm's given by

$$\boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta}(\cdot|B) = Q_{\mu^{B,F},f,\vartheta}\,, \qquad B \in \mathcal{B}\,,$$

is a cp on $(\Omega, \mathcal{A}, \mathcal{B})$ by Remark 2.1 where $\{B \cap f^{-1}(\mathsf{cl}(F_{\Sigma,B}))\colon B \in \mathcal{B}\}$ plays the role of $\mathcal{B}$. In particular, if F equals the convex core then $\boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta} = \boldsymbol{Q}^{\mathcal{B}}_{\mu,f,\vartheta}$.

Theorem 3.3. *If μ is positive and finite on $\mathcal{B}$ then the closure of $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ is the union of the families*

$$\mathfrak{E}^{\mathcal{B},F}_{\mu,f} = \big\{\boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta}\colon\ \vartheta \in \mathsf{cl}(\pi_F(\mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}))) \cap \mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},F,\Sigma f_{\mathcal{B}}})\big\}$$

over the $\mathsf{dom}(\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}})$-accessible faces F of $\mathsf{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$.

Proof. By assumption $\nu = \Sigma f_{\mathcal{B}}\mu_{\mathcal{B}}$ is nonzero and finite, thus [6, Theorem 2] applies to the full standard exponential family $\mathcal{E}_{\nu,\mathsf{id}}$ with $\Xi = \mathsf{dom}(\Lambda_{\nu,\mathsf{id}})$ and implies

$$\mathsf{cl}(\mathcal{E}_{\nu,\mathsf{id}}) = \bigcup \big\{Q_{\nu_F,\mathsf{id},\vartheta}\colon\ \vartheta \in \mathsf{cl}(\pi_F(\Xi)) \cap \mathsf{dom}(\Lambda_{\nu_F,\mathsf{id}})\big\}$$

where the union is over the Ξ-accessible faces F of $\mathsf{cc}(\nu)$ and ν_F denotes the restriction of ν to $\mathsf{cl}(F)$. By Lemma 2.4 *(i)*, $\Lambda_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}$ equals $\Lambda_{\nu,\mathsf{id}}$ so that the above union is over the same family of faces as in the assertion of the theorem.

Lemma 3.2 implies that ν_F is the $\Sigma f_{\mathcal{B}}$-image of the nonzero and finite product measure $\mu_{\mathcal{B},F}$. Hence, $\Lambda_{\nu_F,\mathsf{id}}$ equals $\Lambda_{\mu_{\mathcal{B},F},\Sigma f_{\mathcal{B}}}$ by Lemma 2.4 *(i)*. It follows that in the above union ϑ ranges over the same parameter set as in the assertion of the theorem. Since $Q_{\nu_F,\mathsf{id},\vartheta}$ is the $\Sigma f_{\mathcal{B}}$-image of $Q_{\mu_{\mathcal{B},F},\Sigma f_{\mathcal{B}},\vartheta}$, it is possible to conclude by Lemma 2.4 *(iii)* that

$$\mathsf{cl}(\mathcal{E}_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}}) = \bigcup \big\{Q_{\mu_{\mathcal{B},F},\Sigma f_{\mathcal{B}},\vartheta}\colon\ \vartheta \in \mathsf{cl}(\pi_F(\mathsf{dom}(\Xi))) \cap \mathsf{dom}(\Lambda_{\mu_{\mathcal{B},F},\Sigma f_{\mathcal{B}}})\big\}\,.$$

On account of Lemma 2.2, it suffices to prove that $Q_{\mu_{\mathcal{B},F},\Sigma f_{\mathcal{B}},\vartheta}$ equals $\boldsymbol{\Pi}\boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta}$ but this follows from Lemma 2.3 *(ii)*. □

Corollary 3.4. *If $\Lambda_{\mu^B,f}$ is everywhere finite for all $B \in \mathcal{B}$ then*

$$\mathfrak{E}^{\mathcal{B},F}_{\mu,f} = \big\{\boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta}\colon\ \vartheta \in \mathsf{lin}(F)\big\} \quad \textit{and} \quad \mathsf{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f}) = \bigcup \mathfrak{E}^{\mathcal{B},F}_{\mu,f}$$

where the union is over all faces F of $\mathsf{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$. The mapping $\mathbf{M}_f$ restricts to a bijection between $\mathsf{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$ and $\sum_{B\in\mathcal{B}} \mathsf{cc}(f\mu^B)$.

Proof. The assumption implies that $\mathit{dom}(\Lambda_{\mu_{\mathcal{B},F},\Sigma f_{\mathcal{B}}}) = \mathbb{R}^d$ for all faces F and that all faces are accessible. To prove the second assertion, Lemma 2.3 *(iii)* is applied to $\mu_{\mathcal{B},F}$ in the role of $\mu_{\mathcal{B}}$. Then, $\mathbf{\Pi}$ restricts to a bijection between $\mathfrak{E}^{\mathcal{B},F}_{\mu,f}$ and $\mathcal{E}_{\mu_{\mathcal{B},F},\Sigma f_{\mathcal{B}}}$. By Remark 2.8, $\mathbf{M}_f$ maps $\mathfrak{E}^{\mathcal{B},F}_{\mu,f}$ bijectively onto $\mathit{ri}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B},F})$. This set equals $\mathit{ri}(F)$ by Lemma 3.2. It follows from Theorem 3.3 that $\mathbf{M}_f$ maps $\mathit{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$ bijectively onto the union of $\mathit{ri}(F)$. The union is equal to $\mathit{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$, and thus to the sum of $\mathit{cc}(f\mu^B)$ by Lemma 2.5 *(ii)*. □

Corollary 3.5. *If* $\sum_{B\in\mathcal{B}} \mathit{cc}(f\mu^B)$ *is bounded and locally simplicial then* $\mathbf{M}_f$ *restricts to a homeomorphism between* $\mathit{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$ *and this sum.*

Proof. The boundedness implies that the mapping $P \mapsto \int_{\Omega^{\mathcal{B}}} \Sigma f_{\mathcal{B}} dP$ is continuous on $\mathit{cl}(\mathcal{E}_{\mu_{\mathcal{B}},\Sigma f_{\mathcal{B}}})$. Its inverse is continuous due to the second assumption, see [7, Remark 5.9]. By Lemma 2.2, the assertion follows. □

Example 3.6. Let $(\Omega,\mathcal{A},\mathcal{B})$, μ and f be as in Example 2.10. The segment $F = \{(t,1)\colon 2 \leqslant t \leqslant 4\}$ is a face of the hexagon $\mathit{cc}(\Sigma f_{\mathcal{B}}\mu_{\mathcal{B}})$. Then F_Σ is the square

$$\{((t,0),(0,0),(0,0),(1,0),(1,0),(r,1))\colon 0 \leqslant t,r \leqslant 1\}$$

and $\Sigma f_{\mathcal{B}}\mu_{\mathcal{B},F}$ is the convolution

$$[\delta_{(0,0)}+\delta_{(1,0)}] * \delta_{(0,0)} * \delta_{(0,0)} * \delta_{(1,0)} * \delta_{(1,0)} * [\delta_{(0,1)}+\delta_{(1,1)}] = \delta_{(2,1)} + 2\delta_{(3,1)} + \delta_{(4,1)}\,.$$

The cp $\boldsymbol{P} = \boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta}(\cdot|B)$, $\vartheta = (t,0) \in \mathit{lin}(F)$, from $\mathit{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$ is given by

$$\begin{aligned}
&\boldsymbol{P}(10|00,10) = \boldsymbol{P}(11|01,11) = \tfrac{e^t}{1+e^t}\\
&\boldsymbol{P}(00|00,01) = \boldsymbol{P}(00|00,11) = \boldsymbol{P}(10|10,01) = \boldsymbol{P}(10|10,11) = 1\,.
\end{aligned}$$

The closure of $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ consists of the family itself and 16 families corresponding to all vertices and edges of the hexagon.

4 Discussion

In this section, the space Ω is finite, $\mathcal{A}$ is the algebra 2^Ω of all subsets of Ω, μ is the counting measure on Ω and f maps Ω to $\mathbb{R}^\Omega$ such that $f(\omega)$ is the vector with the ω-th coordinate equal to 1 and the remaining ones to 0. The family $\mathcal{E}_{\mu,f}$ consists of all pm's P on Ω that are positive in the sense $P(\omega) > 0$, $\omega \in \Omega$.

For $B \subseteq \Omega$ the measure $f\mu^B$ is concentrated on the linearly independent set $f(\Omega)$, and hence $\mathit{cc}(f\mu^B)$ is the simplex Δ_B spanned by the set.

Example 4.1. If $\Omega = \{0,1,\dots,m\}$, $m \geqslant 1$, and $\mathcal{B} = \binom{\Omega}{2}$ then $\sum_{B\in\mathcal{B}} \Delta_B$ is the sum of all segments with the endpoints in $f(\Omega)$. This is the polytope known under the name *permutahedron* [16], equivalently defined as the convex hull of all the points $(\rho(0),\rho(1),\dots,\rho(m))$ where ρ is any permutation of Ω. Assume $A_1,\dots,A_k$ is an ordered partition of Ω such that $\omega < \omega'$ for $\omega \in A_i$, $\omega' \in A_j$ and $1 \leqslant i < j \leqslant k$. The convex hull of the points $(\rho(m),\dots,\rho(1),\rho(0))$ where ρ is any permutation of Ω that satisfies $\rho(A_i) = A_i$, $1 \leqslant i \leqslant k$, is a face F of the permutahedron. It is the sum of the faces

$$F_{\Sigma,B} = \begin{cases} \Delta_B\,, & \text{if } B \subseteq A_i \text{ for some } 1 \leqslant i \leqslant k\,,\\ \{f(\omega)\}\,, & \text{otherwise}\,, \end{cases}$$

over $B = \{\omega, \omega'\} \in \binom{\Omega}{2}$ with $\omega < \omega'$. Hence, for $\vartheta = (\vartheta_\omega)_{\omega\in\Omega} \in \mathbb{R}^\Omega$

$$\boldsymbol{Q}^{\mathcal{B},F}_{\mu,f,\vartheta}(\omega|B) = \begin{cases} e^{\vartheta_\omega}/[e^{\vartheta_\omega} + e^{\vartheta_{\omega'}}]\,, & \text{if } B \subseteq A_i \text{ for some } 1 \leqslant i \leqslant k\,, \\ 1\,, & \text{otherwise}\,. \end{cases}$$

Each cp of $\mathsf{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$ has this form up to a permutation.

Remark 4.2. Let $\Omega^*_{\mathcal{B}}$ denote the set of ordered couples $(\omega|B)$ with $\omega \in B \in \mathcal{B}$. For $\mathcal{B} \subseteq \mathcal{A}$ nonempty, a cp $\boldsymbol{P}$ on $(\Omega, \mathcal{A}, \mathcal{B})$ is uniquely given by its nonnegative values $\boldsymbol{P}(\omega|B)$, $(\omega|B) \in \Omega^*_{\mathcal{B}}$. They are constrained by $\sum_{\omega\in B} P(\omega|B) = 1$, $B \in \mathcal{B}$, and

$$\boldsymbol{P}(\omega|C) = \boldsymbol{P}(\omega|B) \cdot \textstyle\sum_{\omega'\in B} \boldsymbol{P}(\omega'|C)\,, \qquad \omega \in B \subseteq C \text{ and } B, C \in \mathcal{B}.$$

By Remark 2.1, $\boldsymbol{P} \in \mathfrak{E}^{\mathcal{B}}_{\mu,f}$ if and only if there exists a positive measure on Ω that generates $\boldsymbol{P}$. It follows from the general results of [4, (6.3), p. 351] that this takes place if and only if all $\boldsymbol{P}(\omega|B)$ are positive and $\boldsymbol{P}$ satisfies the polynomial constraints

$$\textstyle\prod_{i=1}^n \boldsymbol{P}(A_i|B_i) = \prod_{i=1}^n \boldsymbol{P}(A_i|B_{i+1})$$

for $n \geqslant 1$, $B_1, \ldots, B_{n+1} \in \mathcal{B}$ with $B_1 = B_{n+1}$ and $A_i \subseteq B_i \cap B_{i+1}$, $1 \leqslant i \leqslant n$. Here, it can be assumed equivalently that all A_i's are singletons $\{\omega_i\}$. Such a constraint, will be referred to as *Császár* one.

Remark 4.3. It was observed in [11] that Császár constraints correspond to cycles in the bipartite graph $\mathcal{G}_{\mathcal{B}}$ between Ω and $\mathcal{B}$ with the edge from each $B \in \mathcal{B}$ to each of its elements ω. Since the incidence matrix of any bipartite graph is unimodular [15, 19.2] Császár constraints play a distinguished role in the toric ideal induced by the incidence matrix of $\mathcal{G}_{\mathcal{B}}$, see [11, Proposition 3.4].

Lemma 4.4. *A cp $\boldsymbol{P}$ on $(\Omega, 2^\Omega, \mathcal{B})$ satisfies Császár constraints if and only if it extends to a cp $\boldsymbol{P}'$ on $(\Omega, 2^\Omega, 2^\Omega \setminus \{\emptyset\})$, in the sense $\boldsymbol{P}'(\cdot|B) = \boldsymbol{P}(\cdot|B)$, $B \in \mathcal{B}$.*

A proof is omitted; it is based on [4, (5.9), p. 349] that establishes a connection between the constraints and the generation of a cp from a family of measures ordered according to dimension.

Corollary 4.5. *The closure of $\mathfrak{E}^{\mathcal{B}}_{\mu,f}$ consists of all cp's on $(\Omega, 2^\Omega, \mathcal{B})$ that satisfy Császár constraints.*

Example 4.6. In the situation of Example 4.1 with $m \geqslant 2$, for $B = \{\omega, \omega'\}$ with $\omega < \omega'$ let $\boldsymbol{P}(\omega|B) = 1$ and $\boldsymbol{P}(\omega'|B) = 0$ with the exception $\boldsymbol{P}(0|\{0, m\}) = 0$ and $\boldsymbol{P}(m|\{0, m\}) = 1$. Then $\boldsymbol{P}$ is a cp on $(\Omega, \mathcal{A}, \binom{\Omega}{2})$ that violates the Császár constraint with $n = m+1$ and $B_1 = \{0, 1\}$, …, $B_m = \{m-1, m\}, B_n = \{0, m\}$. Thus, $\boldsymbol{P}$ does not belong to $\mathsf{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$.

Remark 4.7. It is not difficult to see that for $\mathcal{B} = \binom{\Omega}{2}$ every $\boldsymbol{P} \in \mathsf{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$ extends to a cp on $(\Omega, 2^\Omega, 2^\Omega \setminus \{\emptyset\})$ uniquely, see the proof of [10, Lemma 4]. In general, it is only a minor technicality not to admit the singletons of Ω in the sets $\mathcal{B}$.

Remark 4.8. In [11], Császár constraints are interpreted as polynomials and are used to define a multiprojective toric variety. The variety lives in the product of the projective spaces of $\mathbb{C}^B$ over $B \in \mathcal{B}$. A point z of this variety is a $\mathcal{B}$-tuple of

points z_B with the projective coordinates $z_{(\omega|B)}$, $\omega \in B$. By [11, Theorem 4.3], the mapping

$$z \mapsto \sum_{B\in\mathcal{B}} \sum_{\omega\in B} f(\omega)\, \frac{|z_{(\omega|B)}|}{\sum_{\omega'\in B} |z_{(\omega'|B)}|}$$

is a bijection between the nonnegative part of the variety and $\sum_{B\in\mathcal{B}} \boldsymbol{\Delta}_B$. (Note that in the original definition of this mapping, denoted by ν, the column $a_{\cdot i|I}$ must be replaced by its projection to the V-coordinates).

The mapping $\mathbf{M}_f$ moves a cp $\boldsymbol{P}$ on $(\Omega, 2^\Omega, \mathcal{B})$ linearly as

$$\boldsymbol{P} \mapsto \sum_{B\in\mathcal{B}} \sum_{\omega\in B} f(\omega)\, \boldsymbol{P}(\omega|B) = \big(\sum_{B\in\mathcal{B}} \boldsymbol{P}(\omega|B)\big)_{\omega\in\Omega}.$$

By Corollary 3.5, $\mathbf{M}_f$ restricts to the homeomorphism between $\mathsf{cl}(\mathfrak{E}^{\mathcal{B}}_{\mu,f})$ and the sum of $\boldsymbol{\Delta}_B$ over $B \in \mathcal{B}$. On account of Corollary 4.5, the closure corresponds to the nonnegative part of the variety from Remark 4.8. Hence, in the setting of this section, the assertion of Corollary 3.5 is equivalent to the statement of [11, Theorem 4.3].

By Corollary 3.5 and Remark 4.7, the family of cp's on $(\Omega, 2^\Omega, \mathcal{B})$ with $\mathcal{B} = \{B \subseteq \Omega\colon |B| \geqslant 2\}$ is homeomorphic to the permutahedron of Example 4.1 via

$$\boldsymbol{P} \mapsto \big(\sum_{\omega'\in\Omega\setminus\{\omega\}} \boldsymbol{P}(\omega|\{\omega,\omega'\})\big)_{\omega\in\Omega}$$

which is the content of [10, Theorem 1].

References

[1] Barndorff-Nielsen, O., *Information and Exponential Families in Statistical Theory.* Wiley, New York, 1978.

[2] Chentsov, N.N., *Statistical Decision Rules and Optimal Inference.* Translations of Mathematical Monographs, AMS, Providence – Rhode Island, 1982 (Russian original: Nauka, Moscow, 1972).

[3] Coletti, G. and Scozzafava, R., *Probabilistic Logic in a Coherent Setting.* Kluwer Academic Publishers, Dordrecht 2002.

[4] Császár, A., Sur la structure des espaces de probabilité conditionnelle. *Acta Math. Acad. Sci. Hung.* **6** (1955) 337–361.

[5] Csiszár, I. and Matúš, F., Convex cores of measures on $\mathbb{R}^d$. *Studia Sci. Math. Hungar.* **38** (2001) 177–190.

[6] Csiszár, I. and Matúš, F., Closures of exponential families. *Annals of Probability* **33** (2005) 582–600.

[7] Csiszár, I. and Matúš, F., Generalized maximum likelihood estimates for exponential families. *Probab. Th. and Related Fields* **141** (2008) 213–246.

[8] de Finetti, B., Sull'impostazione assiomatica del calcolo delle probabilità. *Annali Univ. Trieste* **19** (1949) 3–55.

[9] de Finetti, B., *Probability, Induction and Statistics.* John Wiley & Sons, London, New York, Sydney, Toronto 1972.

[10] Matúš, F., Conditional probabilities and permutahedron. *Annales de l'Institut H. Poincaré, Probabilités et Statistiques* **39** (2003) 687–701.

[11] Morton, J., Relations among conditional probabilities. August 2008, arXiv: 0808.1149v1.

[12] Rényi, A., On a new axiomatic theory of probability. *Acta Math. Acad. Sci. Hung.* **6** (1955) 285–335.

[13] Rényi, A., Sur les espace simples des Probabilités conditionnelles. *Ann. Inst. Henri Poincaré, Probabilités et Statistiques* **1** (1964) 3–21.

[14] Rockafellar, R.T., *Convex Analysis.* Princeton Univ. Press, Princeton 1970.

[15] Schrijver, A., *Theory of Integer and Linear Programming.* John Wiley & Sons, New York, 1998.

[16] Ziegler, G.M., *Lectures on Polytopes.* Springer-Verlag, New York 1995.

Support Sets in Exponential Families and Oriented Matroid Theory

Johannes Rauh
Max Planck Institute
for Mathematics in the Sciences
rauh@mis.mpg.de

Thomas Kahle
Max Planck Institute
for Mathematics in the Sciences
kahle@mis.mpg.de

Nihat Ay
Max Planck Institute
for Mathematics in the Sciences
nay@mis.mpg.de

Abstract

We discuss how to obtain an implicit description of the closure of a discrete exponential family with a finite set of equations derived from an underlying oriented matroid. These equations are similar to the equations used in algebraic statistics, although they need not be polynomial in the general case. This framework allows us to study the possible support sets of an exponential families with the help of oriented matroid theory. In particular, if two exponential families induce the same oriented matroid, then they have the same support sets.

1 Introduction

In this paper we study exponential families, which are well known statistical models with many nice properties. Let $\mathcal{E}$ be an exponential family on a finite set $\mathcal{X}$, and $\overline{\mathcal{E}}$ its closure. We want to describe the set

$$\mathcal{S} := \left\{\operatorname{supp}(P) \subseteq \mathcal{X} : P \in \overline{\mathcal{E}}\right\}. \tag{1}$$

of all possible support sets occurring in $\overline{\mathcal{E}}$.

The problem of determining the possible support sets in an exponential family is a classical problem in statistics. It amounts to describing the boundary of the most basic statistical models. This problem is related to characterizing the marginal polytope, which can be used, for example, to study the existence or non-existence of the MLE [EFRS06]. One can show that computing the support sets of any exponential family is of the same complexity class as NP hard combinatorial problems such as the problem of finding maximal cuts in graphs, since it is known that the class of marginal polytopes contains the so-called cut polytopes (see [KWA09]). This means that there is no corresponding fast algorithm, unless NP = co-NP [DL97]. Nevertheless, considering only certain subclasses of exponential families, the situation may simplify so that explicit statements

about support sets become possible. For instance, one of the authors discusses support sets of small cardinality in hierarchical models, a particular kind of exponential families [Kah10]. In this paper we find a concise characterization of the support sets in general exponential families with the help of oriented matroids. We hope that this will allow for further theoretical results in this direction.

Although slightly hidden, the connection to oriented matroid theory is very natural. The starting point, and another focus of the presentation, is the implicit description of exponential families for discrete random variables inspired by so called Markov bases [GMS06]. It is described in Theorem 4. We study the—not necessarily polynomial—equations that define the closure of the exponential family and relate them to the oriented matroid of the sufficient statistics of the model. In the case of a rational valued sufficient statistics, our observations reduce to the fact that the non-negative real part of a toric variety is described by a circuit ideal. We emphasize how the proof of this fact uses arguments from oriented matroid theory.

This paper is organized as follows. In Section 2 we develop a theory of implicit representations of exponential families which is analogue to and inspired by algebraic statistics [GMS06]. In contrast to the toric case we do not require the sufficient statistics to take integer values and thereby leave the realm of commutative algebra. What remains is the theory of oriented matroids. We discuss how answers to the support set problem look like in the language of oriented matroids and discuss examples coming from cyclic polytopes. These polytopes are well known in combinatorial convexity for their extremal properties, as stated, for instance, in the Upper Bound Theorem. In Section 3 we discuss the basics of the theory of oriented matroids and reformulate statements from Section 2 in this language, making the connection as clear as possible.

2 Exponential families

We assume a finite set $\mathcal{X} := \{1, \ldots, m\}$ and denote $\mathcal{P}(\mathcal{X})$ the open simplex of probability measures with full support on $\mathcal{X}$. The closure of any set $M \subseteq \mathbb{R}^{\mathcal{X}}$, in the standard topology of $\mathbb{R}^n$, is denoted by $\overline{M}$. Any vector $n \in \mathbb{R}^{\mathcal{X}}$ can be decomposed into its positive and negative part $n = n^+ - n^-$ via $n^+(x) := \max(n(x), 0)$ and $n^-(x) := \max(-n(x), 0)$. For any two vectors $n, p \in \mathbb{R}^{\mathcal{X}}$ we define

$$p^n := \prod_{x \in \mathcal{X}} p(x)^{n(x)}, \tag{2}$$

whenever this product is well defined (e.g. when n and p are both non-negative).

Let q be a positive measure on $\mathcal{X}$ with full support, and let $A \in \mathbb{R}^{d \times m}$ be a matrix of width m. We denote a_x, $x \in \mathcal{X}$, the columns of A. Then we have

Definition 1. The *exponential family* associated to the reference measure q and the matrix A is the set of probability measures

$$\mathcal{E}_{q,A} := \left\{ p_\theta \in \mathcal{P}(\mathcal{X}) : p_\theta(x) = \frac{q(x)}{Z_\theta} \exp\left(\theta^T a_x\right), \theta \in \mathbb{R}^d \right\}, \tag{3}$$

where $Z_\theta := \sum_{x \in \mathcal{X}} q(x) \exp\left(\theta^T a_x\right)$ ensures normalization.

If $q(x) = 1$ for all $x \in \mathcal{X}$, i.e. if q is the uniform measure on $\mathcal{X}$, then the corresponding exponential family is abbreviated with $\mathcal{E}_A$.

In the following we always assume that the matrix A has the vector $(1, \dots, 1)$ in its row span. This means that there exists a dual vector $l_1 \in \left(\mathbb{R}^d\right)^*$ which satisfies $l_1(a_x) = 1$ for all $x \in \mathcal{X}$. There is no loss of generality in this assumption as we can always add an additional row $(1, \dots, 1)$ to A without changing the exponential family.

Remark 2. The exponential family depends on A only through its row span $\mathcal{L}$. Different matrices with the same row span lead to different parametrizations of the same exponential family. In the following it will be convenient to fix one parametrization, hence we work with matrices A instead of vector spaces $\mathcal{L}$.

The geometrical structure of the boundary of $\overline{\mathcal{E}}_{q,A}$ is encoded in the polytope of possible values that the map $A\colon \overline{\mathcal{P}}(\mathcal{X}) \to \mathbb{R}^d, x \mapsto Ax$ takes:

Definition 3. The *convex support* of $\mathcal{E}_{q,A}$ is the polytope

$$\operatorname{cs}(\mathcal{E}_{q,A}) := \operatorname{conv}\{a_x : x \in \mathcal{X}\}. \tag{4}$$

In the context of hierarchical models, the convex support is also called *marginal polytope*.

We will see later that the faces of $\operatorname{cs}(\mathcal{E}_{q,A})$ are in a one-to-one correspondence with the different support sets occurring in $\overline{\mathcal{E}}_{q,A}$. Even more is true: The mapping A, restricted to $\overline{\mathcal{E}}_{q,A}$, defines a homeomorphism $\overline{\mathcal{E}}_{q,A} \cong \operatorname{cs}(\mathcal{E}_{q,A})$ which maps every probability measure $p \in \overline{\mathcal{E}}_{q,A}$ into the face corresponding to its support, see for example [BN78]. This homeomorphism is called the *moment map*. One can use the properties of the moment map to prove Theorem 15 using arguments from the theory of oriented matroids. This will be discussed in the next section.

Note that the parametrization in (3) does not extend to the boundary. This is one of the motivations to move on to an implicit description of the exponential family. The next theorem shows how to obtain an implicit description from $\mathcal{E}_{q,A}$ from the kernel of A. This gives a nice "duality" as the parametrization itself is derived from the image of A.

Theorem 4. *A distribution p is an element of the closure of $\mathcal{E}_{q,A}$ if and only if all the equations*

$$p^{n^+} q^{n^-} = p^{n^-} q^{n^+}, \qquad \text{for all } n \in \ker A, \tag{5}$$

hold for p.

Remark 5. This theorem is a direct generalization of Theorem 3.2 in [GMS06]. There only the polynomial equations among (5) are studied under the additional assumption that A has only integer entries. Moreover, only the uniform reference measure was considered. However, the proof of the theorem generalizes without any major problem. Actually, the proof of our theorem here needs one step less, since we don't need to show the reduction to the polynomial equations. The different flavor of the results will be made more precise in Remark 13 later.

Our proof closely follows [GMS06]. In our presentation of the proof we want to explicitly point out how matroid-type arguments are used, the first example being Lemma 7.

Before giving the proof of Theorem 4 we first state a couple of auxiliary results which are of independent interest. The matrix A and derived objects are fixed for the rest of the considerations. A face of a polytope P is the intersection of the polytope with an affine hyperplane H, such that all $x \in P$ with $x \notin H$ lie on one side of the hyperplane. Faces of maximal dimension are called *facets*. It is a fundamental result that every polytope can equivalently be described as a compact set defined by finitely many inequalities (i.e. facets), see [Zie94].

In particular we are interested in the face structure of $\mathrm{cs}(\mathcal{E}_{q,A})$. Since we assumed that all columns of A lie in the affine hyperplane $l_1 = 1$, we can replace every affine hyperplane H by an equivalent central hyperplane (which passes through the origin). This motivates the following

Definition 6. Let $\{a_x : x \in \mathcal{X}\}$ be the vertex set of a polytope. A set $F \subseteq \mathcal{X}$ is called *facial* if there exists a vector $c \in \mathbb{R}^d$ such that

$$c^T a_y = 0 \quad \forall y \in F, \qquad\qquad c^T a_z \geq 1 \quad \forall z \notin F. \tag{6}$$

Lemma 7. *Fix a matrix $A = (a_x)_{x\in\mathcal{X}} \in \mathbb{R}^{d\times m}$ and a nonempty subset $F \subseteq \mathcal{X}$. Then we have:*

- *If F is facial then no non-zero non-negative linear combination of the a_x, $x \notin F$, can be written as linear combination of the a_x, $x \in F$.*
- *F is facial if and only if for any $u \in \ker A$:*

$$\mathrm{supp}(u^+) \subseteq F \Leftrightarrow \mathrm{supp}(u^-) \subseteq F. \tag{7}$$

- *If p is a solution to (5), then $\mathrm{supp}(p)$ is facial.*

Proof. For the first statement, assume to the contrary that we can find $\alpha(x) \geq 0$ and $\beta(x)$ not all zero such that $u = \sum_{x\notin F} \alpha(x)a_x = \sum_{x\in F} \beta(x)a_x$, and let c be as in (6). We have

$$0 \leq \sum_{i\notin F} \alpha_i \leq \sum_{i\notin F} \alpha_i c^T a_i = c^T \left(\sum_{i\notin F} \alpha_i a_i\right) = c^T \left(\sum_{i\in F} \beta_i a_i\right) = 0,$$

whence $\alpha_i = 0$ for all $i \notin F$. This also proves the first direction of the second statement.

The opposite direction is a bit more complicated and uses Farkas' Lemma (see for example [Zie94]): *Let $B \in \mathbb{R}^{l\times d}$, and $z \in \mathbb{R}^l$. Either there exists a point in the polyhedron $\{x : Bx \leq z\}$, or there exists a non-negative vector $y \in \mathbb{R}^l_{\geq}$ with $y^T B = 0$ and $y^T z < 0$, but not both.* Assume that $F \subsetneq \mathcal{X}$ is nonempty and satisfies (7) for all $u \in \ker A$. Let B be the $(|F| + m) \times d$ matrix with rows $\left\{a_x^T : x \in F\right\}$, $\left\{-a_x^T : x \in F\right\}$, $\left\{-a_x^T, x \notin F\right\}$, and z be the vector which has entries zero in the first $2\,|F|$ components and entries -1 in the last $m - |F|$. Then a solution to $Bx \leq z$ provides a facial vector. Thus it remains to show that each non-negative $y = (y^{(1)}, y^{(2)}, y^{(3)})^T$, decomposed according to the rows of B, with $y^T B = 0$ satisfies $y^T z \geq 0$. Assume that the columns of A are ordered such that the columns with indices $x \in F$ come first. Then $y^{(3)}$ must be zero as otherwise $(y^{(2)} - y^{(1)}, y^{(3)})^T \in \ker A$ would violate (7) by non-negativity of y. But then $y^T z = 0$ trivially.

The last statement follows immediately from the second statement. □

Now we are ready for the proof of Theorem 4.

Proof of Theorem 4. The first thing to note is that it is enough to prove the theorem when $q(x) = 1$ for all x. To see this note that $p \in \overline{\mathcal{E}}_A$ if and only if $\lambda qp \in \overline{\mathcal{E}}_{q,A}$, where $\lambda > 0$ is a normalizing constant, which does not appear in equations (5) since they are homogeneous.

Denote Z_A the set of solutions of (5). We first show that $\mathcal{E}_A$ satisfies the equations defining Z_A. We plug in the parametrization to find

$$p^u = \prod_{x\in\mathcal{X}} p(x)^{u(x)} = \prod_{x\in\mathcal{X}} \left(\mathrm{e}^{\theta^T a_x}\right)^{u(x)} = \prod_{x\in\mathcal{X}} \mathrm{e}^{\theta(x)(Au)(x)} = \prod_{x\in\mathcal{X}} \mathrm{e}^{\theta(x)(Av)(x)} = p^v. \tag{8}$$

Thus $\mathcal{E}_A \subseteq Z_A$, and also $\overline{\mathcal{E}}_A \subseteq \overline{Z}_A = Z_A$.

Next, let $p \in Z_A \setminus \mathcal{E}_A$. We construct a sequence p_μ in $\mathcal{E}_A$ that converges to p as $\mu \to -\infty$.

Consider the following system of equations in variables $d = (d_1, \ldots, d_n)$:

$$d^T a_x = \log p(x) \quad \text{for all } x \in \operatorname{supp}(p). \tag{9}$$

We claim that this linear system has a solution. Otherwise we can find numbers $v(x)$, $x \in F$, such that $\sum_x v(x) \log p(x) \neq 0$ and $\sum_x v(x) a_x = 0$. This leads to the contradiction $p^{v^+} \neq p^{v^-}$.

Fix a vector $c \in \mathbb{R}^d$ with property (6) and for any $\mu \in \mathbb{R}$ define

$$p_{(\mu)} := p_{\mu c + d} = \left(\mathrm{e}^{\mu c^T a_1} \mathrm{e}^{d^T a_1}, \ldots, \mathrm{e}^{\mu c^T a_m} \mathrm{e}^{d^T a_m}\right) \in \mathcal{E}_A.$$

By (6) it is clear that $\lim_{\mu\to-\infty} p_{(\mu)} = p$. This proves the theorem. □

We now see that the last statement of Lemma 7 can be generalized [GMS06, Lemma A.2]:

Proposition 8. *The following are equivalent for any set $F \subseteq \mathcal{X}$:*

1. *F is facial.*
2. *The uniform distribution $\frac{1}{|F|}\mathbb{1}_{\{F\}}$ of F lies in $\overline{\mathcal{E}}_A$.*
3. *There is a vector with support F in $\overline{\mathcal{E}}_A$.*

According to Theorem 4, in order test whether p is an element of the closure of $\mathcal{E}_{q,A}$, we have to test all the equations (5). The next theorem shows that it is actually enough to check finitely many equations. For this, we need the following notion from matroid theory: A *circuit vector* of a matrix A is a nonzero vector $n \in \mathbb{R}^m$ corresponding to a linear dependency $\sum_x n(x) a_x$ with inclusion minimal support, i.e if $n' \in \mathbb{R}^m$ satisfies $\operatorname{supp}(n') \subseteq \operatorname{supp}(n)$, then $n' = \lambda n$ for some $\lambda \in \mathbb{R}$. Equivalently, n is an element of $\ker A$ with inclusion minimal support.

A *circuit* is the support set of a circuit vector. The minimality condition implies that the circuit determines its corresponding circuit vectors up to a multiple. A *circuit basis* C contains one circuit vector for every circuit.[1]

[1] It is easy to see that a circuit basis of $\ker A$ spans $\ker A$. However, in general the circuit vectors are not linearly independent.

If we replace n by a nonzero multiple of n then equation (5) is replaced by an equation which is equivalent over the non-negative reals. This means that all systems of equations corresponding to any circuit basis C are equivalent.

Theorem 9. *Let $\mathcal{E}_{q,A}$ be an exponential family. Then $\overline{\mathcal{E}}_{q,A}$ equals the set of all probability distributions that satisfy*

$$p^{c^+} q^{c^-} = p^{c^-} q^{c^+} \text{ for all } c \in C, \tag{10}$$

where C is a circuit basis of A.

The proof is based on the following two lemmas:

Lemma 10. *For every vector $n \in \ker A$ there exists a sign-consistent circuit vector $c \in \ker A$, i.e. if $c(x) \neq 0$ then $\operatorname{sgn} c(x) = \operatorname{sgn} n(x)$ for all $x \in \mathcal{X}$.*

Proof. Let c be a vector with inclusion-minimal support which is sign-consistent with n and satisfies $\operatorname{supp}(c) \subseteq \operatorname{supp}(n)$. If c is not a circuit, then there exists a circuit c' with $\operatorname{supp}(c') \subseteq \operatorname{supp}(c)$. Using a suitable linear combination $c + \alpha c'$, $\alpha \in \mathbb{R}$, we can obtain a contradiction to the minimality of c. □

Lemma 11. *Every vector $n \in \ker A$ is a finite sign-consistent sum of circuit vectors $n = \sum_{i=1}^{r} c_i$, i.e. if $c_i(x) \neq 0$ then $\operatorname{sgn} c_i(x) = \operatorname{sgn} n(x)$ for all $x \in \mathcal{X}$.*

Proof. Use induction on the size of $\operatorname{supp}(n)$. In the induction step, use a sign-consistent circuit, as in the last lemma, to reduce the support. □

Proof of Theorem 9. Again, we can assume that $q(x) = 1$ for all $x \in \mathcal{X}$. By Theorem 4 it suffices to show: If $p \in \mathbb{R}^{\mathcal{X}}$ satisfies (10), then it also satisfies $p^{n^+} = p^{n^-}$ for all $n \in \ker A$. Write $n = \sum_{i=1}^{r} c_i$ as a sign-consistent sum of circuits c_i, as in the last lemma. Without loss of generality we can assume $c_i \in C$ for all i. Then $n^+ = \sum_{i=1}^{r} c_i^+$ and $n^- = \sum_{i=1}^{r} c_i^-$. Hence p satisfies

$$p^{n^+} - p^{n^-} = p^{\sum_{i=1}^{r-1} c_i^+} \left(p^{c_1^+} - p^{c_1^-} \right) + \left(p^{\sum_{i=1}^{r-1} c_i^+} - p^{\sum_{i=1}^{r-1} c_i^-} \right) p^{c_1^-}, \tag{11}$$

so the theorem follows easily by induction. □

The theorem implies that a finite number of equations is sufficient to describe $\overline{\mathcal{E}}_{q,A}$. The number of equations that are necessary is bounded from above by the number of different support sets occurring in C.

Example 12. Consider the following sufficient statistics:

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -\alpha & 1 & 0 & 0 \end{pmatrix}, \tag{12}$$

where $\alpha \notin \{0, 1\}$ is arbitrary. The kernel is then spanned by

$$v_1 = (1, \alpha, -1, -\alpha)^T \text{ and } v_2 = (1, \alpha, -\alpha, -1)^T. \tag{13}$$

These two generators correspond to the two relations

$$p(1)p(2)^\alpha = p(3)p(4)^\alpha, \text{ and } p(1)p(2)^\alpha = p(3)^\alpha p(4). \tag{14}$$

It follows immediately that

$$p(3)p(4)^\alpha = p(3)^\alpha p(4). \tag{15}$$

If $p(3)p(4)$ is not zero, then we conclude $p(3) = p(4)$. However, on the boundary this does not follow from equations (14): Possible solutions to these equations are given by

$$p_a = (0, a, 0, 1-a) \text{ for } 0 \leq a < 1. \tag{16}$$

However, p_a does not lie in the closure of the exponential family $\overline{\mathcal{E}}_A$, since all members of $\mathcal{E}_A$ do satisfy $p(3) = p(4)$.

A circuit basis of A is given by the following vectors:

$$(0,0,1,-1)^T \qquad p(3) = p(4), \tag{17a}$$

$$(1,\alpha,0,-1-\alpha)^T \qquad p(1)p(2)^\alpha = p(4)^{1+\alpha}, \tag{17b}$$

$$(1,\alpha,-1-\alpha,0)^T \qquad p(1)p(2)^\alpha = p(3)^{1+\alpha}. \tag{17c}$$

Remark 13 (Relation to algebraic statistics). In the particular case where the vector space $\ker A$ has a basis with integer components (for example, if A itself has only integer entries), every circuit is proportional to a circuit with integer components. In this case the corresponding equations (5) are polynomial, and the theorem implies that $\overline{\mathcal{E}}_A$ is the non-negative real part of a *projective variety*, i.e. the solution set of homogeneous polynomials. If we want to use the tools of commutative algebra and algebraic geometry, then it turns out that circuits are not the right object to consider: For example, proportional circuits only yield equivalent equations if we consider them over the non-negative reals, but we may obtain a different solution set if we allow negative real solutions or complex solutions, which may greatly increase the running time of many algorithms of computational commutative algebra. Hence, if we want to use algebraic tools, it is best to work with a *Markov basis*, which can be defined as a finite set of kernel vectors such that the solution set over $\mathbb{C}$ of the corresponding equations equals the *Zariski closure* of $\overline{\mathcal{E}}$, i.e. the smallest variety containing $\overline{\mathcal{E}}$.[2] In this algebraic setting, Theorem 4 remains valid if we replace "closure" by "Zariski closure" and $\ker A$ by the integer kernel $\ker_{\mathbb{Z}} A$. This fact was first noted in [DS98].

In the algebraic case one can also look at the *ideal* (see [CLO08]) generated by all polynomial equations induced by integer valued circuit vectors. This ideal is called the *circuit ideal*. By what was said above this ideal is in general smaller than the associated *toric ideal*, which contains the polynomial equations induced by all integer valued kernel vectors. Circuit ideals have been studied already in the seminal paper [ES96]. For further results illuminating their nice relations to polyhedral geometry we refer to [BJT07].

Finding a Markov basis is in general a non-trivial task, see [HM09]. It seems to be much easier to compute the circuits of a matrix. However, a minimal Markov basis is usually much smaller than a circuit basis, and thus it is easier to handle (but cf. the next remark). For experiments in this direction we recommend the open source software package `4ti2` [4ti2] which can compute circuits as well as Markov bases.

[2] It turns out that it is not so easy to find an example of a Markov basis which does not consists of circuits. In [AT03], S. Aoki and A. Takemura give a model and a Markov basis element which is not a circuit. Interestingly, the full Markov basis of this model is not known.

Remark 14. Using arguments from matroid theory the number of circuits can be shown to be less or equal than $\binom{m}{r+2}$, where $m = |\mathcal{X}|$ is the size of the state space and r is the dimension of the exponential family $\mathcal{E}_{q,A}$, see [DSL04]. This gives us an upper bound on the number of implicit equations which is necessary to describe $\overline{\mathcal{E}}_{q,A}$. Note that $\binom{m}{r+2}$ is usually much larger than the codimension $m-r-1$ of $\mathcal{E}_{q,A}$ in the probability simplex. In contrast to this, if we only want to find an implicit description of all probability distributions of $\mathcal{E}_{q,A}$, which have full support, then $m-r-1$ equations are enough: We can test $p \in \mathcal{E}_{q,A}$ by checking whether $\log(p/q)$ lies in the column span of A. This amounts to checking whether $\log(p/q)$ is orthogonal to $\ker A$, which is equivalent to $m-r-1$ equations, once we have chosen a basis of $\ker A$.

It turns out that even in the boundary the number of equations can be further reduced: In general we do not need all circuits for the implicit description of $\overline{\mathcal{E}}_{q,A}$. For instance, in Example 12, the equations 17b and 17c are equivalent given 17a, i.e. we only need two of the three circuits to describe $\overline{\mathcal{E}}_{q,A}$. Unfortunately we do not know how to find a minimal subset of circuits that characterizes the closure of the exponential family. Of course, in the algebraic case discussed in the previous remark this question is equivalent to determining a minimal generating set of the circuit ideal among the circuits.

Now we focus on the following problem: Given a set $S \subseteq \mathcal{X}$, is there a probability distribution $p \in \overline{\mathcal{E}}_A$ satisfying $\operatorname{supp}(p) = S$? In other words, we want to characterize the set

$$\mathcal{S}(q, A) := \{\operatorname{supp}(p) : p \in \overline{\mathcal{E}}_{q,A}\} \subseteq 2^{\mathcal{X}}. \tag{18}$$

Proposition 8 gives the following characterization: A nonempty set $S \subseteq \mathcal{X}$ is the support set of some distribution $p \in \overline{\mathcal{E}}_{q,A}$ if and only if the following holds for all circuit vectors $n \in \ker A$:

- $\operatorname{supp}(n^+) \subseteq S$ if and only if $\operatorname{supp}(n^-) \subseteq S$.

Obviously, this condition does not depend on the circuits themselves, but only on the supports of their positive and negative part. In order to formalize this observation, consider the map

$$\operatorname{sgn}\colon n \mapsto (\operatorname{supp}(n^+), \operatorname{supp}(n^-)),$$

which associates to each vector a pair of disjoint subsets of $\mathcal{X}$. Such a pair of disjoint subsets shall be called a *signed subset* of $\mathcal{X}$ in the following. Alternatively, signed subsets (A, B) can also be represented as sign vectors $X \in \{-1, 0, +1\}^{\mathcal{X}}$, where

$$X(x) = \begin{cases} +1, & \text{if } x \in A, \\ -1, & \text{if } x \in B, \\ 0, & \text{else.} \end{cases} \tag{19}$$

In this representation, sgn corresponds to the usual signum mapping extended to vectors. As a slight abuse of notation, we don't make a difference between these two representations in the following.

The signed subset $\operatorname{sgn}(c)$ corresponding to a circuit $c \in \ker A$ shall be called an *oriented circuit*. The set of all oriented circuits is denoted by

$$\mathcal{C}(A) := \pm \operatorname{sgn}(C) = \{\operatorname{sgn}(c) : c \in C \text{ or } c \in -C\}, \tag{20}$$

where C is a circuit basis of A.

We immediately have the following

Theorem 15. *Let S be a nonempty subset of $\mathcal{X}$. Then $S \in \mathcal{S}$ if and only if the following holds for all signed circuits $(A, B) \in \mathcal{C}(A)$:*

$$A \subseteq S \quad \Leftrightarrow \quad B \subseteq S. \tag{21}$$

Corollary 16. *If two matrices A_1, A_2 satisfy $\mathcal{C}(A_1) = \mathcal{C}(A_2)$ then the possible support sets of the corresponding exponential families $\mathcal{E}_{q_1,A_1}$ and $\mathcal{E}_{q_2,A_2}$ coincide.*

According to remark 14, Theorem 15 gives us up to $\binom{m}{r+2}$ conditions on the support. Usually, some of these conditions are redundant, but it is not easy to see a priori, which conditions are essential. Of course, a necessary condition for a subset S of $\mathcal{X}$ to be a support set of a distribution contained in $\overline{\mathcal{E}}_A$ is condition (21) restricted to pairs from a subset $\mathcal{H} \subseteq \mathcal{C}(A)$. For example, one can take $\mathcal{H} := \operatorname{sgn}(B)$, where B is a finite subset of $\ker A$, such as a basis.

Example 17. Let's continue Example 12. From the circuits we deduce the following implications:

$$p(3) \neq 0 \quad \Longleftrightarrow \quad p(4) \neq 0, \tag{22a}$$

$$p(1) \neq 0 \text{ and } p(2) \neq 0 \quad \Longleftrightarrow \quad p(4) \neq 0, \tag{22b}$$

$$p(1) \neq 0 \text{ and } p(2) \neq 0 \quad \Longleftrightarrow \quad p(3) \neq 0. \tag{22c}$$

Again, as above, the last two implications are equivalent given the first.

From this it follows easily that the possible support sets in this example are $\{1\}$, $\{2\}$ and $\{1, 2, 3, 4\}$. From the spanning set (13) we only obtain the implication

$$p(1) \neq 0 \text{ and } p(2) \neq 0 \quad \Longleftrightarrow \quad p(3) \neq 0 \text{ and } p(4) \neq 0. \tag{23}$$

We conclude this section with two examples where a complete characterization of the face lattice of the convex support and thus of the possible supports is easily achievable.

Example 18 (Supports in the binary no-n-way interaction model). Consider the binary hierarchical model [KWA09] whose simplicial complex is the boundary of an n simplex. If $n = 3$, this model is called the no-3-way interaction model and its Markov bases have been recognized to be arbitrarily complicated [LO06], so we cannot hope to find an easy description of the oriented circuits. However, if we restrict ourselves to binary variables $x = (x_i)_{i=1}^n \in \mathcal{X} := \{0, 1\}^n$, the structure is very simple. In this case the exponential family is of dimension $2^n - 2$, i.e. of codimension 1 in the simplex, so $\ker A$ is one dimensional. It is spanned by the "parity function":

$$e_{[n]}(x) := \begin{cases} -1 & \text{if } \sum_{i=1}^n x_i \text{ is odd,} \\ 1 & \text{otherwise.} \end{cases} \tag{24}$$

Using Theorem 15 we can easily describe the face lattice of the marginal polytope (i.e. convex support) $P^{(n-1)}$: A set $\mathcal{Y} \subsetneq \{0, 1\}^n$ is a support set if and only if it does not contain all configurations with even parity, or all configurations with odd parity. It follows that $P^{(n-1)}$ is *neighborly*, i.e. the convex hull of any

$\lfloor \frac{\dim(P^{(n-1)})}{2} \rfloor = 2^{n-1} - 1$ vertices is a face of the polytope. To see this, note that no set of cardinality less than 2^{n-1} can contain all configurations with even or odd parity. We can easily count the support sets by counting the non-faces of the corresponding marginal polytope, i.e. all sets $\mathcal{Y}$ that contain either the configurations with even parity, or the configurations with odd parity. Let s_k be the number of support sets of cardinality of k, i.e. the number of faces with k vertices. It is given by:

$$s_k = \binom{2^n}{k} - 2\binom{2^{n-1}}{k - 2^{n-1}}, \tag{25}$$

where $\binom{m}{l} = 0$ if $l < 0$. Since this polytope has only one affine dependency (24) which includes all the vertices, we see that it is *simplicial*, i.e. all its faces are simplices. It follows that f_k, the number of k-dimensional faces, is given by $f_k = s_{k-1}$.

Altogether we have determined the face lattice of the polytope, which means that we know the "combinatorial type" of the polytope. It turns out that the face lattice of $P^{(n-1)}$ is isomorphic to the face lattice of the $(2^n - 2)$-dimensional *cyclic polytope* with 2^n vertices.

Next, we take a closer look at cyclic polytopes. Define the *moment curve* in $\mathbb{R}^d$ by

$$\boldsymbol{x} : \mathbb{R} \to \mathbb{R}^d, \qquad t \mapsto \boldsymbol{x}(t) := \left(t, t^2, \cdots, t^d\right)^T. \tag{26}$$

The *d-dimensional cyclic polytope with n vertices* is

$$C(d, n) := \operatorname{conv}\left\{\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_n)\right\}, \tag{27}$$

the convex hull of $n > d$ distinct points $(t_1 < t_2 < \ldots < t_n)$ on the moment curve. The face lattice of a cyclic polytope can easily be described using *Gale's evenness condition*, see [Zie94]. The cyclic polytope is simplicial and neighborly, i.e. the convex hull of any $\lfloor \frac{d}{2} \rfloor$ vertices is a face of $C(n, d)$, but even better, one has

Theorem 19 (Upper Bound Theorem)**.** *If P is a d-dimensional polytope with $n = f_0$ vertices, then for every k it has at most as many k-dimensional faces as the cyclic polytope $C(d, n)$:*

$$f_k(P) \leq f_k(C(d, n)), \quad k = 0, \ldots, d. \tag{28}$$

If equality holds for some k with $\lfloor \frac{d}{2} \rfloor \leq k \leq d$ then P is neighborly.

Theorem 19 was conjectured by Motzkin in 1957 and its proof has a long and complicated history. The final result is due to McMullen [McM70].

The Upper Bound Theorem shows that the exponential families constructed above have the largest number of support sets among all exponential families with the same dimension and the same number of vertices. Finally, we consider a cyclic polytope of dimension two which also gives an interesting exponential family, answering the question for the exponential family of smallest dimension containing all the vertices of the probability simplex. The construction is due to [MA04].

Example 20. Let $\mathcal{X} = \{1, \ldots, m\}$ and consider the matrix A, whose columns are the points on the 2-dimensional moment curve, augmented with row $(1, \ldots, 1)$:

$$A := \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 2 & 3 & \ldots & m \\ 1 & 4 & 9 & \ldots & m^2 \end{pmatrix}. \tag{29}$$

This matrix defines a two-dimensional exponential family. To approximate an arbitrary extreme point δ_j of the probability simplex, consider the parameter vector $\theta = (j^2, -2j, 1)^T$, giving rise to probability measures $p_{\beta\theta} = \frac{1}{Z}\exp(-\beta\theta^T A)$. Since $\theta^T A_i = (i-j)^2$, we get that $\lim_{\beta\to\infty} p_{\beta\theta} = \delta_j$.

Summarizing we see that cyclic polytopes, owing to their extremal properties, have something to offer not only for convex geometry, but also for statistics.

3 Relations to Oriented Matroids

In this section the results from the previous section are related to the theory of oriented matroids. The proofs in this section are only sketched, since the main results of this work have already been proved directly. We refer to chapters 1 to 3 of [BVS+93] for a more detailed introduction to oriented matroids.

Let E be a finite set and $\mathcal{C}$ a non-empty collection of signed subsets of E (see the previous section). For every signed set $X = (X^+, X^-)$ of E we let $\underline{X} := X^+ \cup X^-$ denote the *support* of X. Furthermore, the *opposite signed set* is $-X = (X^-, X^+)$. Then the pair $(E, \mathcal{C})$ is called an *oriented matroid* if the following conditions are satisfied:

(C1) $\mathcal{C} = -\mathcal{C}$,

(C2) for all $X, Y \in \mathcal{C}$, if $\underline{X} \subseteq \underline{Y}$, then $X = Y$ or $X = -Y$, (*incomparability*)

(C3) for all $X, Y \in \mathcal{C}$, $X \neq -Y$, and $e \in X^+ \cap Y^-$ there is a $Z \in \mathcal{C}$ such that $Z^+ \subseteq (X^+ \cup Y^+) \setminus \{e\}$ and $Z^- \subseteq (X^- \cup Y^-) \setminus \{e\}$. (*weak elimination*)

In this case each element of $\mathcal{C}$ is called a *signed circuit*.

Note that to every oriented matroid $(E, \mathcal{C})$ we have an associated unoriented matroid (E, C), called the *underlying matroid*, where

$$C = \{X^+ \cup X^- = \operatorname{supp}(X) : X \in \mathcal{C}\} \tag{30}$$

is the set of *circuits* of (E, C). In this way oriented matroids can be considered as ordinary matroids endowed with an additional structure, namely a *circuit orientation* which assigns two opposite signed circuits $\pm X \in \mathcal{C}$ to every circuit $X \in C$.

The most important example of an oriented matroid here is the oriented matroid of a matrix $A \subseteq \mathbb{R}^{d\times m}$. In this case let $E = \mathcal{X} = \{1, \ldots, m\}$, and let

$$\mathcal{C} = \big\{(\operatorname{supp}(n^+), \operatorname{supp}(n^-) : n \in \ker A \text{ has inclusion minimal support.}\big\}. \tag{31}$$

This example is so important that oriented matroids which arise in this way are given a name: An oriented matroid is called *realizable* if it is induced by some matrix A.[3]

[3] Note that this definition depends, in fact, only on the kernel of A, compare Remark 2

The only axiom which is not trivially fulfilled for this example is **(C3)**. However, if we drop the minimality condition and let $\mathcal{V} = \{(\mathrm{supp}(n^+), \mathrm{supp}(n^-) : n \in \ker A\}$, then it is easy to see that $\mathcal{V}$ satisfies **(C3)**. Thus $(E, \mathcal{C})$ satisfies **(C3)** by the following proposition:

Proposition 21. *Let $\mathcal{V}$ be a nonempty collection of signed subsets of E satisfying* **(C1)** *and* **(C3)**. *Write* $\mathrm{Min}(\mathcal{V})$ *for the minimal elements of $\mathcal{V}$ (with respect to inclusion of supports). Then*

1. *for any $X \in \mathcal{V}$ there is $Y \in \mathrm{Min}(\mathcal{V})$ such that $Y^+ \subseteq X^+$ and $Y^- \subseteq X^-$.*
2. $\mathrm{Min}(\mathcal{V})$ *is the set of circuits of an oriented matroid.*

Proof. [BVS+93], proposition 3.2.4. □

This illustrates how **(C2)** corresponds to the minimality condition. It is possible to define oriented matroids without this minimality condition using the following construction:

For two signed subsets X, Y of E define the *composition* of X and Y as

$$(X \circ Y)^+ := X^+ \cup (Y^+ \setminus X^-), \qquad (X \circ Y)^- := X^- \cup (Y^- \setminus X^+). \tag{32}$$

Note that this operation is associative but not commutative in general. A composition $X \circ Y$ is *conformal* if X and Y are *sign-consistent*, i.e. $X^+ \cap Y^- = \emptyset = X^- \cap Y^+$.

An *o.m. vector* of an oriented matroid is any composition of an arbitrary number of circuits.[4] The set of o.m. vectors shall be denoted by $\mathcal{V}$. If the oriented matroid comes from a matrix A, then $\mathcal{V}$ equals the set $\mathcal{V}$ from above.

The above proposition implies easily that an oriented matroid can be defined as a pair $(E, \mathcal{V})$, where $\mathcal{V}$ is a collection of signed subsets satisfying **(C1)**, **(C3)** and

(V0) $\emptyset \in \mathcal{V}$,

(V2) for all $X, Y \in \mathcal{V}$ we have $X \circ Y \in \mathcal{V}$,

Note that in the realizable case linear combinations of vectors correspond to composition of their sign vectors in the following sense:

$$\mathrm{sgn}(n + \epsilon n') = \mathrm{sgn}(n) \circ \mathrm{sgn}(n'), \qquad \text{for } \epsilon > 0 \text{ small enough.} \tag{33}$$

Now Lemmas 10 and 11 correspond to the following two lemmas

Lemma 10'. *For every o.m. vector Y there exists a sign-consistent signed circuit X such that $\underline{X} \subseteq \underline{Y}$.*

Lemma 11'. *Any o.m. vector is a conformal composition of circuits.*

To every matrix A we can associate a polytope which was called convex support in the last section. Many properties of this polytope can be translated into the language of oriented matroids. This yields constructions which also make sense, if the oriented matroid is not realizable. In order to make this

[4] In [BVS+93], o.m. vectors are simply called vectors. The name "o.m. vector" has been proposed by F. Matúš to avoid confusion.

more precise, we need the notion of the dual oriented matroid. The general construction of the dual of an oriented matroid is beyond the scope of this work. Here, we only state the definition for realizable oriented matroids.

In the following we assume that the matrix A has the constant vector $(1,\dots,1)$ in its rowspace. This means that all the column vectors a_x lie in a hyperplane $l_1 = 1$. In the general case, this can always be achieved by adding another dimension. Technically we require that the face lattice of the polytope spanned by the columns of A is combinatorially equivalent to the face lattice of the cone over the columns. See also the remarks before Definition 6.

For every dual vector $l \in (\mathbb{R}^d)^*$ let $N_l^+ := \{x \in \mathcal{X} : l(a_x) > 0\}$ and $N_l^- := \{x \in \mathcal{X} : l(a_x) < 0\}$. This way we can associate a signed subset $\operatorname{sgn}^*(l) := (N_l^+, N_l^-)$ with l. The signed subset $\operatorname{sgn}^*(l)$ is called a *covector*. Let $\mathcal{L}$ be the set of all covectors. If the signed subset (N_l^+, N_l^-) has minimal support (i.e. "many" vectors a_x lie on the hyperplane $l = 0$), then l is called a *cocircuit vector*, and $\operatorname{sgn}^*(l)$ is called a *signed cocircuit*. The collection of all signed cocircuits shall be denoted by $\mathcal{C}^*$.

Lemma 22. *Let $(E, \mathcal{C})$ be an oriented matroid induced by a matrix A. Then $(E, \mathcal{C}^*)$ is an oriented matroid, called the* dual oriented matroid.

Proof. See section 3.4 of [BVS+93]. □

Note that the faces of the polytope correspond to hyperplanes such that all vertices lie on one side of this hyperplane, compare Definition 6. Thus the faces of the polytope are in a one-to-one relation with the positive covectors, i.e. the covectors $X = (X^+, X^-)$ such that $X^- = \emptyset$. The face lattice of the polytope can be reconstructed by partially ordering the positive covectors by inclusion of their supports; however, the relation needs to be inverted: Covectors with small support correspond to faces which contain many vertices. The empty face (which is induced, for example, by the dual vector l_1 which defines the hyperplane containing all a_x) corresponds to the covector $T := (\mathcal{X}, \emptyset)$.

We can apply these remarks to all abstract oriented matroids such that $T = (\mathcal{X}, \emptyset)$ is a covector. Such an oriented matroid is usually called *acyclic*. Thus a face of an acyclic oriented matroid is any positive covector. A *vertex* is a maximal positive covector X in $\mathcal{L} \setminus \{T\}$, i.e. if $\underline{X} \subseteq \underline{Y}$ for some positive covector $Y \in \mathcal{L} \setminus \{X\}$, then $Y = T$.

In this setting we have the following result, which clearly corresponds to the second statement of 7:

Proposition 23 (Las Vergnas)**.** *Let $(E, \mathcal{C})$ be an acyclic oriented matroid. For any subset $F \subseteq E$ the following are equivalent:*

- *F is a face of the oriented matroid.*
- *For every signed circuit $X \in \mathcal{C}$, if $X^+ \subseteq F$ then $X^- \subseteq F$.*

Proof. The proof of Proposition 9.1.2 in [BVS+93] applies (note that the statement of Proposition 9.1.2 includes an additional assumption which is never used in the proof). □

With the help of the moment map defined in the previous section, this proposition can be used to easily derive Theorem 15: By the properties of the moment map, every face of the convex support corresponds to a possible support

set of an exponential family, and the proposition links this to the signed circuits of the corresponding oriented matroid.

Finally, Corollary 16 can be rewritten as

Corollary 16'. *The possible support sets of two exponential families coincide if they have the same oriented matroids.*

Unfortunately, this correspondence is not one-to-one: Different oriented matroids can yield the same face lattice, i.e. combinatorially equivalent polytopes. A simple example is given by a regular and a non-regular octahedron as described in [Zie94]. The special case has a name: an oriented matroid is *rigid*, if its positive covectors (i.e. its face lattice) determine all covectors (i.e. the whole oriented matroid). Still, Corollary 16' implies that the instruments of the theory of oriented matroids should suffice to describe the support sets of an exponential family.

Remark 24 (Importance of Duality). There are mainly two reasons why the theory of oriented matroids (as well as the theory of ordinary matroids) is considered important. First, it yields an abstract framework which allows to describe a multitude of different combinatorial questions in a unified manner. This, of course, does not in itself lead to any new theorem. The second reason is that the theory provides the important tool of matroid duality.

It turns out that the dual of a realizable matroid is again realizable: If A is a matrix representing an oriented matroid $(E, \mathcal{C})$, then any matrix A^* such that the rows of A^* span the orthogonal complement of the row span of A represents the oriented matroid $(E, \mathcal{C}^*)$.

To motivate the importance of this construction we sketch its implications for the case that the oriented matroid comes from a polytope. In this case the duality is known under the name *Gale transform* [Zie94, Chapter 6]. A d-dimensional polytope with N vertices can be represented by N vectors in $\mathbb{R}^{d+1}$ lying in a hyperplane. These vectors form a $(d+1) \times N$-matrix A. Now we can find an $(N-d-1) \times N$-matrix A^* as above, so the dual matroid is represented by a configuration of N vectors in $\mathbb{R}^{N-d-1}$. This means that this construction allows us to obtain a lowdimensional image of a highdimensional polytope, as long as the number of vertices is not much larger than the dimension. This method has been used for example in [Stu88] in order to construct polytopes with quite unintuitive properties, leading to the rejection of some conjectures. Furthermore, oriented matroid duality makes it possible to classify polytopes with "few vertices" by classifying vector configurations.

The notion of dimension generalizes to arbitrary oriented matroids (and ordinary matroids). In the general setting one usually talks about the *rank* of a matroid, which is defined as the maximal cardinality of a subset $E \subseteq F$ such that E contains no support of a signed circuit. In this sense duality exchanges examples of high rank and low rank, where "high" and "low" is relative to $|E|$.

Acknowledgement

The authors want to thank Fero Matúš and Bastian Steudel for many discussions. Nihat Ay has been supported by the Santa Fe Institute. Thomas Kahle has been supported by the Volkswagen Foundation.

References

[4ti2] 4ti2 team, *4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces*, available at www.4ti2.de.

[AT03] Satoshi Aoki and Akimichi Takemura, *The list of indispensable moves of the unique minimal markov basis for 3x4xk and 4x4x4 contingency tables with fixed two-dimensional marginals*, METR 2003-38, University of Tokyo, Japan, 2003.

[BJT07] Tristram Bogarta, Anders N. Jensen, and Rekha R. Thomas, *The circuit ideal of a vector configuration*, Journal of Algebra **309** (2007), no. 2, 518–542.

[BN78] Ole E. Barndorff-Nielsen, *Information and exponential families in statistical theory*, Wiley, 1978.

[BVS+93] A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G. Ziegler, *Oriented matroids*, Encyclopedia of Mathematics and its Applications, Cambridge, 1993.

[CLO08] David A. Cox, John B. Little, and Don O'Shea, *Ideals, varieties, and algorithms: An introduction to computational algebraic geometry and commutative algebra*, Springer, 2008.

[DL97] Michel M. Deza and Monique Laurent, *Geometry of cuts and metrics*, Algorithms and Combinatorics, Springer, 1997.

[DS98] Persi Diaconis and Bernd Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Annals of Statistics **26** (1998), 363–397.

[DSL04] György Dósa, István Szalkai, and Claude Laflamme, *The maximum and minimum number of circuits and bases of matroids*, Pure Mathematics and Applications **15** (2004), no. 4, 383–392.

[EFRS06] Nicholas Eriksson, Stephen E. Fienberg, Alessandro Rinaldo, and Seth Sullivant, *Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models.*, Journal of Symbolic Computation (2006), no. 41, 222–233, Special issue on Computational Algebraic Statistics.

[ES96] David Eisenbud and Bernd Sturmfels, *Binomial ideals*, Duke Mathematical Journal **84** (1996), no. 1, 1–45.

[GMS06] Dan Geiger, Christopher Meek, and Bernd Sturmfels, *On the toric algebra of graphical models*, The Annals of Statistics **34** (2006), no. 5, 1463–1492.

[HM09] Raymond Hemmecke and Peter Malkin, *Computing generating sets of lattice ideals*, Journal of Symbolic Computation (2009), accepted.

[Kah10] Thomas Kahle, *Neighborliness of marginal polytopes*, Contributions to Algebra and Geometry (2010), accepted, arXiv:0809.0786.

[KWA09] Thomas Kahle, Walter Wenzel, and Nihat Ay, *Hierarchical models, marginal polytopes, and linear codes*, Kybernetika **45** (2009), no. 2, 189–208.

[LO06] Jesús A. De Loera and Shmuel Onn, *Markov bases of three-way tables are arbitrarily complicated*, Journal of Symbolic Computation **41** (2006), 173–181.

[MA04] František Matúš and Nihat Ay, *On maximization of the information divergence from an exponential family*, Kybernetika **41** (2004), 731–746.

[McM70] Peter McMullen, *The maximum number of faces of a convex polytope*, Mathematika (1970), 179–184.

[Stu88] Bernd Sturmfels, *Some applications of affine gale diagrams to polytopes with few vertices*, SIAM J. Discrete Mathematics **1** (1988), 121–133.

[Zie94] Günter M. Ziegler, *Lectures on polytopes*, GTM, vol. 152, Springer Verlag, 1994.

Mixtures of Polynomials in Hybrid Bayesian Networks with Deterministic Variables

Prakash P. Shenoy
School of Business
University of Kansas
pshenoy@ku.edu

James C. West
School of Business
University of Kansas
cully@ku.edu

Abstract

The main goal of this paper is to describe inference in hybrid Bayesian networks (BNs) using mixtures of polynomials (MOP) approximations of probability density functions (PDFs). Hybrid BNs contain a mix of discrete, continuous, and conditionally deterministic random variables. The conditionals for continuous variables are typically described by conditional PDFs. A major hurdle in making inference in hybrid BNs is marginalization of continuous variables, which involves integrating combinations of conditional PDFs. In this paper, we suggest the use of MOP approximations of PDFs, which are similar in spirit to using mixtures of truncated exponentials (MTEs) approximations. MOP functions can be easily integrated, and are closed under combination and marginalization. This enables us to propagate MOP potentials in the extended Shenoy-Shafer architecture for inference in hybrid BNs that can include deterministic variables. MOP approximations have several advantages over MTE approximations of PDFs. They are easier to find, even for multi-dimensional conditional PDFs, and are applicable for a larger class of deterministic functions in hybrid BNs.

1 Introduction

Bayesian networks (BNs) and influence diagrams (IDs) were invented in the mid 80s (see e.g., [17], [7]) to represent and reason with large multivariate discrete probability models and decision problems, respectively. Several efficient algorithms exist to compute exact marginals of posterior distributions for discrete BNs (see e.g., [11], [23], and [9]) and to solve discrete IDs exactly (see e.g., [16], [20], [21], and [8]).

The state of the art exact algorithm for mixtures of Gaussians hybrid BNs is the Lauritzen-Jensen algorithm [10]. This requires the conditional PDFs of continuous variables to be conditional linear Gaussians, and that discrete variables do not have continuous parents. Marginals of multivariate normal distributions can be found easily without the need for integration. The disadvantages are that in the inference process, continuous variables have to be marginalized before discrete ones. In some problems, this restriction can lead to large cliques

[12].

If a BN has discrete variables with continuous parents, Murphy [15] uses a variational approach to approximate the product of the potentials associated with a discrete variable and its parents with a conditional linear Gaussian. [13] uses a numerical integration technique called Gaussian quadrature to approximate non-conditional linear Gaussian distributions with conditional linear Gaussians, and this same technique can be used to approximate the product of potentials associated with a discrete variable and its continuous parents. Murphy's and Lerner's approach is then embedded in the Lauritzen-Jensen algorithm [10] to solve the resulting mixtures of Gaussians BN.

Shenoy [22] proposes approximating non-conditional linear Gaussian distributions by mixtures of Gaussians using a nonlinear optimization technique, and using arc reversals to ensure discrete variables do not have continuous parents. The resulting mixture of Gaussians BN is then solved using Lauritzen-Jensen algorithm [10].

[14] proposes approximating PDFs by mixtures of truncated exponentials (MTE), which are easy to integrate in closed form. Since the family of mixtures of truncated exponentials is closed under combination and marginalization, the Shenoy-Shafer architecture [23] can be used to solve a MTE BN. [4] proposes using a non-linear optimization technique for finding MTE approximations for several commonly used one-dimensional distributions. [2, 3] extend this approach to BNs with linear and non-linear deterministic variables. In the latter case, they approximate non-linear deterministic functions by piecewise linear ones.

In this paper, we propose using mixtures of polynomials (MOP) approximations of PDFs. Mixtures of polynomials are widely used in many domains including computer graphics, font design, approximation theory, and numerical analysis. They were first studied by Schoenberg [18]. When the MOP functions are continuous, they are referred to as polynomial splines [19]. The use of splines to approximate PDFs was initially suggested by [5]. For our purposes, continuity is not an essential requirement, and we will restrict our analysis to piecewise polynomial approximations of PDFs.

Using MOP is similar in spirit to using MTEs. MOP functions can be easily integrated, and they are closed under combination and marginalization. Thus, the extended Shenoy-Shafer architecture [25] can be used to make inferences in BN with deterministic variables. However, there are several advantages of MOP functions over MTEs.

First, we can find MOP approximations of differentiable PDFs easily by using the Taylor series approximations. Finding MTE approximations as suggested by [4] necessitates solving non-linear optimization problems, which is not as easy a task as it involves navigating among local optimal solutions.

Second, for the case of conditional PDFs with several parents, finding a good MTE approximation can be extremely difficult as it involves solving a non-linear optimization problem in high-dimensional space for each piece. The Taylor series expansion can also be used for finding MOP approximations of conditional PDFs. In [24], we describe a MOP approximation for a 2-dimensional CLG distribution.

Third, if a hybrid BN contains deterministic functions, then the MTE approach can be used directly only for linear deterministic functions. By directly, we mean without approximating a non-linear deterministic function by a piece-

wise linear one. This is because the MTE functions are not closed under transformations needed for non-linear deterministic functions. MOP functions are closed under a larger family of deterministic functions including linear functions and quotients [24]. This enables propagation in a bigger family of hybrid BNs than is possible using MTEs.

An outline of the remainder of the paper is as follows. In Section 2, we define MOP functions and describe how one can find MOP approximations with illustration for the univariate normal distribution. In Section 3, we solve a small example designed to demonstrate the feasibility of using MOP approximations with a non-differentiable deterministic function. Finally, in Section 4, we end with a summary and discussion of some of the challenges associated with MOP approximations.

2 Mixtures of Polynomials Approximations

In this section, we describe MOP functions and some methods for finding MOP approximations of PDFs. We illustrate our method for the normal distribution. In [24], we also describe MOP approximations of the PDFs of the chi-square distribution, and the conditional linear Gaussian distribution in two dimensions.

2.1 MOP Functions

A one-dimensional function $f : \mathcal{R} \rightarrow \mathcal{R}$ is said to be a *mixture of polynomials* (MOP) function if it is a piecewise function of the form:

$$f(x) = \begin{cases} a_{0i} + a_{1i}x + a_{2i}x^2 + \cdots + a_{ni}x^n & \text{for } x \in A_i, i = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

where $A_1, \ldots, A_k$ are disjoint intervals in $\mathcal{R}$ that do not depend on x, and $a_{0i}, \ldots, a_{ni}$ are constants for all i. We will say that f is a k-piece (ignoring the 0 piece), and n-degree (assuming $a_{ni} \neq 0$ for some i) MOP function.

The main motivation for defining MOP functions is that such functions are easy to integrate in closed form, and that they are closed under multiplication and integration. They are also closed under differentiation and addition.

An m-dimensional function $f : \mathcal{R}^m \rightarrow \mathcal{R}$ is said to be a MOP function if:

$$f(x_1, \ldots, x_m) = f_1(x_1) \cdot f_2(x_2) \cdots f_m(x_m) \tag{2.2}$$

where each $f_i(x_i)$ is a one-dimensional MOP function as defined in Equation (2.1). If $f_i(x_i)$ is a k_i-piece, n_i-degree MOP function, then f is a $(k_1 \cdots k_m)$-piece, $(n_1 + \ldots + n_m)$-degree MOP function. Therefore it is important to keep the number of pieces and degrees to a minimum.

2.2 Finding MOP Approximations of PDFs

Consider the univariate standard normal PDF $\phi(z) = (1/\sqrt{2\pi})e^{-z^2/2}$. A 1-piece, 28-degree, MOP approximation $\phi_{1p}(z)$ of $\phi(z)$ in the interval $(-3, 3)$ is as follows:

$$\phi_{1p}(z) = \begin{cases} c^{-1}(1 - z^2/2 + z^4/8 - \ldots + z^{28}/1428329123020800) & \text{if } -3 < z < 3, \\ 0 & \text{otherwise} \end{cases}$$

where $c^{-1} \approx 0.4$. This MOP approximation was found using the Taylor series expansion of $e^{-z^2/2}$, at $z = 0$, to degree 28, restricting it to the region $(-3, 3)$, verifying that $\phi_{1p}(z) \geq 0$ in the region $(-3, 3)$, and normalizing it with constant c so that $\int \phi_{1p}(z)dz = 1$ (whenever the limits of integration are not specified, the entire range $(-\infty, \infty)$ is to be understood). We will denote these operations by writing:

$$\phi_{1p}(z) = \begin{cases} TSeries[e^{-z^2/2}, z = 0, d = 28] & \text{if } -3 < z < 3 \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

We can verify that $\phi_{1p}(z) \geq 0$ as follows. First, we plot the unnormalized MOP approximation, denoted by, say, $\phi_u(z)$. From the graph, we identify approximately the regions where $\phi_u(z)$ could possibly be negative. Then starting from a point in each these regions, we compute the local minimum of $\phi_u(z)$ using, e.g., gradient descent. Since MOP functions are easily differentiable, the gradients can be easily found. If $\phi_u(z) \geq 0$ at all the local minimums, then we have verified that $\phi_{1p}(z) \geq 0$. If $\phi_u(z) < 0$ at a local minimum, then we need to either increase the degree of the polynomial approximation, or increase the number of pieces, or both.

We have some very small coefficients in the MOP approximation. Rounding these off to a certain number of decimal places could cause numerical instability. Therefore, it is important to keep the coefficients in their rational form.

A graph of the MOP approximation $\phi_{1p}(z)$ overlaid on the actual PDF $\phi(z)$ is shown in Figure 1 and it shows that there are not many differences between the two functions in the interval $(-3, 3)$. The main difference is that ϕ_{1p} is restricted to $(-3, 3)$, whereas ϕ is not. The mean of ϕ_{1p} is 0, and its variance ≈ 0.976. Most of the error in the variance is due to the restriction of the distribution to the interval $(-3, 3)$. If we restrict the standard normal density ϕ function to the interval $(-3, 3)$, renormalize it so that it is a PDF, then its variance ≈ 0.973.

In some examples, working with a 28-degree polynomial may not be tractable. In this case, we can include more pieces to reduce the degree of the polynomial. For example, a 6-piece, 3-degree MOP approximation of $\phi(z)$ is as follows:

$$\phi_{6p}(z) = \begin{cases} TSeries[e^{-z^2/2}, z = -5/2, d = 3] & \text{if } -3 < z < -2, \\ TSeries[e^{-z^2/2}, z = -3/2, d = 3] & \text{if } -2 \leq z < -1, \\ TSeries[e^{-z^2/2}, z = -1/2, d = 3] & \text{if } -1 \leq z < 0, \\ TSeries[e^{-z^2/2}, z = 1/2, d = 3] & \text{if } 0 \leq z < 1, \\ TSeries[e^{-z^2/2}, z = 3/2, d = 3] & \text{if } 1 \leq z < 2, \\ TSeries[e^{-z^2/2}, z = 5/2, d = 3] & \text{if } 2 \leq z < 3, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Notice that ϕ_{6p} is discontinuous at the end points of the intervals. Also, $E(\phi_{6p}) = 0$, and $V(\phi_{6p}) \approx 0.974$. The variance of ϕ_{6p} is closer to the variance of the truncated normal (≈ 0.973) than ϕ_{1p}.

In some examples, for reasons of precision, we may wish to work with a larger interval than $(-3, 3)$ for the standard normal. For example, an 8-piece,

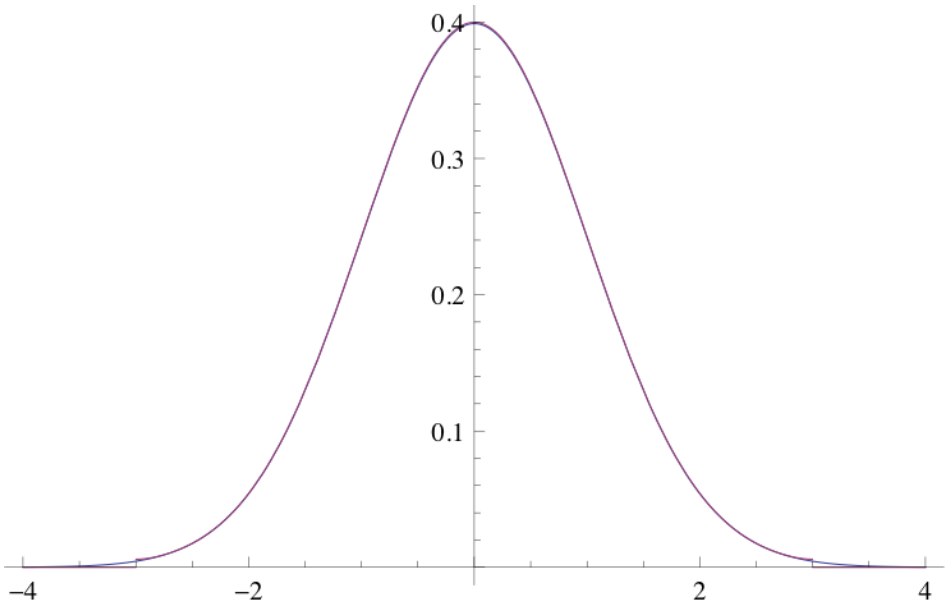


Figure 1: A graph of $\phi_{1p}(z)$ overlaid on $\phi(z)$

4-degree MOP approximation of ϕ in the interval $(-4, 4)$ is as follows:

$$\phi_{8p}(z) = \begin{cases} TSeries[e^{-z^2/2}, z = -7/2, d = 4] & \text{if } -4 < z < -3, \\ TSeries[e^{-z^2/2}, z = -5/2, d = 3] & \text{if } -3 \leq z < -2, \\ TSeries[e^{-z^2/2}, z = -3/2, d = 3] & \text{if } -2 \leq z < -1, \\ TSeries[e^{-z^2/2}, z = -1/2, d = 3] & \text{if } -1 \leq z < 0, \\ TSeries[e^{-z^2/2}, z = 1/2, d = 3] & \text{if } 0 \leq z < 1, \\ TSeries[e^{-z^2/2}, z = 3/2, d = 3] & \text{if } 1 \leq z < 2, \\ TSeries[e^{-z^2/2}, z = 5/2, d = 3] & \text{if } 2 \leq z < 3, \\ TSeries[e^{-z^2/2}, z = 7/2, d = 4] & \text{if } 3 \leq z < 4, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Notice that the degrees of the first and the eighth pieces are 4 to avoid $\phi_{8p}(z) < 0$. $E(\phi_{8p}(z)) = 0$, and $V(\phi_{8p}(z)) \approx 0.99985$. Due to the larger interval, the variance is closer to 1 than the variance for ϕ_{6p}. If we truncate the PDF of the standard normal to the region $(-4, 4)$ and renormalize it, then its variance is ≈ 0.99893.

To find a MOP approximation of the PDF of the $N(\mu, \sigma^2)$ distribution, where μ and $\sigma > 0$ are constants, we exploit the fact that MOP functions are invariant under linear transformations. Thus, if $f(x)$ is a MOP function, then $f(ax + b)$ is also a MOP function. If $Z \sim N(0, 1)$, its PDF is approximated by a MOP function $\phi_p(z)$, and $X = \sigma Z + \mu$, then $X \sim N(\mu, \sigma^2)$, and a MOP approximation of the PDF of X is given by $\xi(x) = (1/\sigma)\phi_p((x - \mu)/\sigma)$.

3 An Example

In this section, we illustrate the use of MOP functions for solving a small hybrid Bayesian network (BN) with a deterministic variable. We use the extended Shenoy-Shafer architecture described in [25]. In [24], we solve more hybrid BNs with deterministic variables including the quotient and the product deterministic functions.

Consider a BN as shown in Figure 2. X and Y are continuous variables and W is deterministic with a non-differentiable function of X and Y, $W = \max\{X, Y\}$.

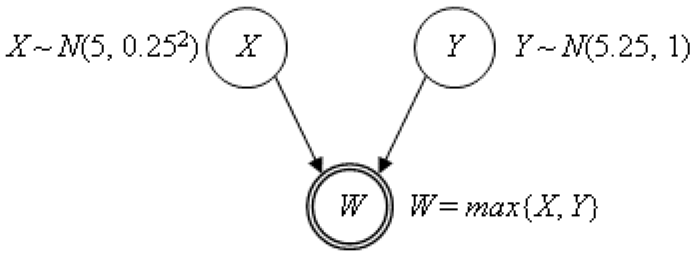


Figure 2: A BN with a max deterministic function

The conditional associated with W is represented by the Dirac potential $\omega(x, y, w) = \delta(w - \max\{x, y\})$, where δ is a Dirac delta function [6]. To compute the marginal PDF of W, we need to evaluate the integral

$$f_W(w) = \int f_X(x)(\int f_Y(y)\delta(w - \max\{x, y\})dy)dx \tag{3.1}$$

where $f_W(w)$, $f_X(x)$, and $f_Y(y)$ are the marginal PDFs of W, X, and Y, respectively. Since the deterministic function is not differentiable, the integrals in Equation (3.1) cannot be evaluated as written.

One solution to finding the marginal PDF of W is to use theory of order statistics. Let $F_W(w)$, $F_X(x)$, and $F_Y(y)$ denote the marginal cumulative distribution functions (CDFs) of W, X, and Y, respectively. Then:

$$F_W(w) = P(W \leq w) = P(X \leq w, Y \leq w) = F_X(w)F_Y(w). \tag{3.2}$$

Differentiating both sides of Equation (3.2) with respect to w, we have:

$$f_W(w) = f_X(w)F_Y(w) + F_X(w)f_Y(w). \tag{3.3}$$

In our example, X and Y have normal PDFs, which does not have a closed form CDF. However, using MOP approximations of the normal PDF, we can easily compute a closed form expression for the CDFs, which will remain MOP functions. Then, using Equation (3.3), we will have a closed-form MOP approximation for the PDF of W. Assuming we start with the 8-piece, 4-degree MOP approximation ϕ_{8p} of $N(0, 1)$ on the interval $(-4, 4)$ as described in Equation (2.5), we can find MOP approximations of the PDFs of $N(5, 0.25^2)$ and $N(5.25, 1)$ as discussed in Section 2 as follows:

$$\xi(x) = 4\phi_{8p}(4(x - 5)), \tag{3.4}$$

$$\psi(y) = \phi_{8p}(y - 5.25). \tag{3.5}$$

Next we find the MOP approximations of the CDFs of X and Y, and then the MOP approximation of the PDF of W using Equation (3.3). A graph of the MOP approximation of $f_W(w)$ is shown in Figure 3.

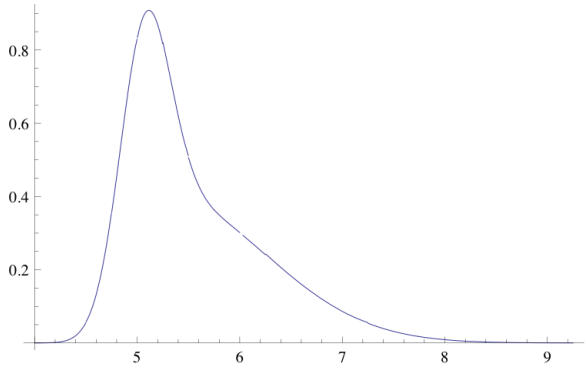


Figure 3: A graph of the MOP approximation of the PDF of W

The mean and variance of the MOP approximation of f_W are computed as 5.5484 and 0.4574. [1] provides formulae for exact computation of the mean and variance of the max of two normals as follows:

$$E(W) = E(X)F_Z(b) + E(Y)F_Z(-b) + af_Z(b), \tag{3.6}$$

$$E(W^2) = (E(X)^2 + V(X))F_Z(b) + (E(Y)^2 + V(Y))F_Z(-b) + (E(X) + E(Y))af_z(b), \tag{3.7}$$

where $a^2 = V(X) + V(Y) - 2C(X,Y)$, $b = (E(X) - E(Y))/a$, and f_Z and F_Z are the PDF and CDF of $N(0,1)$, respectively.

In our example, $E(X) = 5$, $E(Y) = 5.25$, $V(X) = 0.25^2$, $V(Y) = 1$, $C(X,Y) = 0$. Thus, $E(W) \approx 5.5483$, and $V(W) \approx 0.4576$. The mean and variance of the MOP approximation of W are accurate to three decimal places. Unfortunately, the reasoning behind this computation of the marginal of W is not included in inference in BNs.

To obtain the marginal of W using BN inference, we convert the max function to a differentiable function as follows: $\max\{X,Y\} = X$ if $X \geq Y$, and $= Y$ if $X < Y$. We include a discrete variable A with two states, a and na, where a indicates that $X \geq Y$, and make it a parent of W. The revised BN is shown in Figure 4.

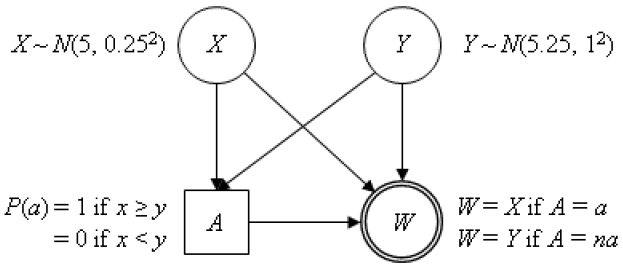


Figure 4: The revised BN for the max deterministic function

Starting with the BN in Figure 4, the marginal of W can be computed using the extended Shenoy-Shafer architecture [25]. We start with mixed potentials as follows:

$$\mu_X(x) = (1, \xi(x)); \tag{3.8}$$

$$\mu_y(y) = (1, \psi(y)); \tag{3.9}$$

$$\mu_A(a, x, y) = (H(x-y), 1), \mu_A(na, x, y) = (1 - H(x-y), 1); \tag{3.10}$$

$$\mu_W(a, x, y, w) = (1, \delta(w-x)), \mu_W(na, x, y, w) = (1, \delta(w-y)). \tag{3.11}$$

In Equation (3.10), $H(.)$ is the *Heaviside* function such that $H(x) = 1$ if $x \geq 0$, and $= 0$ otherwise. The Heaviside function is a MOP function.

To find the marginal of W, we sequentially delete X, Y, and A. To delete X, first we combine μ_X, μ_A, and μ_W, and then marginalize X from the combination:

$$(\mu_X \otimes \mu_A \otimes \mu_W)(a, x, y, w) = (H(x-y), \xi(x)\delta(w-x)), \tag{3.12}$$

$$(\mu_X \otimes \mu_A \otimes \mu_W)(na, x, y, w) = (1 - H(x-y), \xi(x)\delta(w-y)); \tag{3.13}$$

$$\begin{aligned}(\mu_X \otimes \mu_A \otimes \mu_W)^{-X}(a, y, w) &= (1, \int H(x-y)\xi(x)\delta(w-x))dx) \\ &= (1, H(w-y)\xi(w)), \end{aligned} \tag{3.14}$$

$$\begin{aligned}(\mu_X \otimes \mu_A \otimes \mu_W)^{-X}(na, y, w) &= (1, \delta(w-y) \int (1 - H(x-y))\xi(x)dx) \\ &= (1, \delta(w-y)\theta(y)); \end{aligned} \tag{3.15}$$

where $\theta(y) = \int (1 - H(x-y))\xi(x)dx$.

Next, we delete Y. To do so, we combine $(\mu_X \otimes \mu_A \otimes \mu_W)^{-X}$ and μ_Y, and then marginalize Y:

$$((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)(a, y, w) = (1, H(w-y)\xi(w)\psi(y)), \tag{3.16}$$

$$((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)(na, y, w) = (1, \delta(w-y)\theta(y)\psi(y)); \tag{3.17}$$

$$\begin{aligned}((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y}(a, w) &= (1, \xi(w) \int H(w-y)\psi(y)dy) \\ &= (1, \xi(w)\rho(w)), \end{aligned} \tag{3.18}$$

$$((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y}(na, w) = (1, \theta(w)\psi(w)); \tag{3.19}$$

where $\rho(w) = \int H(w-y)\psi(y)dy$.

Finally, we delete A by marginalizing A from $((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y}$:

$$\begin{aligned}(((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y})^{-A}(w) &= (1, \xi(w)\rho(w) + \theta(w)\psi(w)) \\ &= (1, \omega(w)); \end{aligned} \tag{3.20}$$

where $\omega(w) = \xi(w)\rho(w) + \theta(w)\psi(w)$. $\omega(w)$ is a MOP approximation of $f_W(w)$. Notice that

$$\rho(w) = \int H(w-y)\psi(y)dy = F_Y(w), \text{and} \tag{3.21}$$

$$\theta(w) = \int (1 - H(x-y))\xi(x)dx = 1 - P(X > w) = F_X(w), \tag{3.22}$$

and therefore, $\omega(w) = \xi(w)\rho(w) + \theta(w)\psi(w)$ is a MOP approximation of $f_X(w)F_Y(w) + F_X(w)f_Y(w)$. We get exactly the same results as those obtained by using theory of order statistics but using BN inference.

4 Summary and Discussion

The biggest problem associated with inference in hybrid BNs is the integration involved in marginalization of continuous variables. As a remedy, we have proposed MOP approximations for PDFs in the same spirit as MTE approximations [14]. Like MTE functions, MOP functions are easy to integrate, and are closed under combination and marginalization. This allows propagation of MOP potentials using the extended Shenoy-Shafer architecture [25].

MOP approximations have several advantages over MTE approximations of PDFs. First, they are easy to find using the Taylor series expansion of differentiable functions. Second, finding MOP approximations of multi-dimensional conditional PDFs is also relatively straightforward using the multi-dimensional Taylor series expansion. Third, MOP approximations are closed for a larger family of deterministic functions including the quotient functions. Beyond these observations, a formal empirical comparison of MOP vs. MTE approximations is an issue that needs further study.

Some issues associated with MOP approximations that need to be investigated are as follows. There is a tradeoff between the number of pieces and the degree of the polynomial. More pieces mean smaller intervals and consequently smaller degrees. Assuming the goal is to find marginals most efficiently, what is the optimal number of pieces/degrees?

Another challenge is to describe the effect of pieces/terms on the errors in the moments of marginals. It appears that most of the errors in the moments are caused by truncating the domain of variables to some finite intervals. Thus, it may be possible to decide on what intervals should be used if we wish to compute marginals within some prescribed error bounds for the moments of the marginal of variable of interest.

High degree MOP approximations lead to very small coefficients that need to be kept in rational form. This may decrease the efficiency of computation, and may limit the size of BN models that can be solved. One solution here is to use more pieces, which lowers the degrees of the MOP approximations.

MOP approximations are not closed for many classes of deterministic functions such as products and exponentiation. If we can expand the class of MOP functions to include positive and negative rational exponents and maintain the properties of MOP functions—easily integrable, closed under combination and marginalization—then we can solve hybrid BNs with a larger class of deterministic functions.

References

[1] C. E. Clark. The greatest of a finite set of random variables. *Operations Research*, 9(2):145–162, 1961.

[2] B. R. Cobb and P. P. Shenoy. Hybrid Bayesian networks with linear deterministic variables. In F. Bacchus and T. Jaakkola, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 21st Conference*, pages 136–144, Corvallis, OR, 2005. AUAI Press.

[3] B. R. Cobb and P. P. Shenoy. Nonlinear deterministic relationships in Bayesian networks. In L. Godo, editor, *Symbolic and Quantitative Ap-*

proaches to Reasoning with Uncertainty: 8th European Conference, ECSQARU 2005, Lecture Notes in Artificial Intelligence 3571, pages 27–38, Berlin, 2005. Springer.

[4] B. R. Cobb, P. P. Shenoy, and R. Rumi. Approximating probability density functions in hybrid Bayesian networks with mixtures of truncated exponentials. *Statistics & Computing*, 16(3):293–308, 2006.

[5] R. M. Curds. *Propagation techniques in probabilistic expert systems.* PhD thesis, Department of Statistical Science, University College London, London, UK, 1997.

[6] P. A. M. Dirac. The physical interpretation of the quantum dynamics. *Proceedings of the Royal Society of London, Series A*, 113(765):621–641, 1927.

[7] R. A. Howard and J. E. Matheson. Influence diagrams. In R. A. Howard and J. E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 719–762. Strategic Decisions Group, 1984.

[8] F. Jensen, F. V. Jensen, and S. L. Dittmer. From influence diagrams to junction trees. In R. L. de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 10th Conference*, pages 367–373, San Francisco, CA, 1994. Morgan Kaufmann.

[9] F. V. Jensen, S. L. Lauritzen, and K. G. Olesen. Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly*, 4:269–282, 1990.

[10] S. L. Lauritzen and F. Jensen. Stable local computation with conditional Gaussian distributions. *Statitistics and Computing*, 11:191–203, 2001.

[11] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[12] U. Lerner and R. Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In J. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 17th Conference*, pages 310–318, San Francisco, CA, 2001. Morgan Kaufmann.

[13] U. Lerner, E. Segal, and D. Koller. Exact inference in networks with discrete children of continuous parents. In J. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 17th Conference*, pages 319–328, San Francisco, CA, 2001. Morgan Kaufmann.

[14] S. Moral, R. Rumí, and A. Salmerón. Mixtures of truncated exponentials in hybrid Bayesian networks. In S. Benferhat and P. Besnard, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 6th European Conference, ECSQARU-2001*, Lecture Notes in Artificial Intelligence 2143, pages 156–167, Berlin, 2001. Springer.

[15] K. P. Murphy. A variational approximation for Bayesian networks with discrete and continuous latent variables. In K. Laskey and H. Prade, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 15th Conference*, pages 457–466, San Francisco, CA, 1999. Morgan Kaufmann.

[16] S. M. Olmsted. *On representing and solving decision problems*. PhD thesis, Department of Engineering-Economic Systems, Stanford University, Stanford, CA., 1983.

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

[18] I. J. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics*, 4(45-99):112–141, 1946.

[19] L. L. Schumaker. *Spline Functions: Basic Theory*. Cambridge University Press, Cambridge, UK, third edition, 2007.

[20] R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

[21] P. P. Shenoy. Valuation-based systems for Bayesian decision analysis. *Operations Research*, 40(3):463–484, 1992.

[22] P. P. Shenoy. Inference in hybrid Bayesian networks using mixtures of Gaussians. In R. Dechter and T. Richardson, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 22nd Conference*, pages 428–436, Corvallis, OR, 2006. AUAI Press.

[23] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. In R. D. Shachter, T. Levitt, J. F. Lemmer, and L. N. Kanal, editors, *Uncertainty in Artificial Intelligence 4*, pages 169–198. North-Holland, 1990.

[24] P. P. Shenoy and J. C. West. Inference in hybrid Bayesian networks using mixtures of polynomials. Working Paper 321, University of Kansas School of Business, Lawrence, KS, May 2009.

[25] P. P. Shenoy and J. C. West. Inference in hybrid Bayesian networks with deterministic variables. In C. Sossai and G. Chemello, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty–10th ECSQARU*, Lecture Notes in Artificial Intelligence 5590, pages 46–58, Berlin, 2009. Springer-Verlag.

Comparison of Probabilistic and Possibilistic Approaches to Modelling of Economic Uncertainty

Hans Schjær-Jacobsen
Copenhagen University College of Engineering
2750 Ballerup, Denmark
hsj@ihk.dk

Abstract

Recent ex post research into mega-projects has disclosed that cost usually is severely underestimated and benefits severely overestimated. It is thus proposed to improve the practical methods for taking economic risk and uncertainty more thoroughly into account. For that purpose we first compare probabilistic and possibilistic approaches and then combine them into a unified concept of imprecise stochastic variables. Numerical examples demonstrate the potential of the methods considered.

1 Introduction

The traditional approach to representation of uncertainty in economics is that of **probabilities** [1]. An uncertain parameter may be represented by a probability distribution reflecting either the objective nature of the parameter or the decision maker's subjective belief. The probabilistic approach is particularly suited for representing the statistical nature or variability of a parameter. In the general case where the actual economic problem under consideration is modelled by a function of uncertain parameters, Monte Carlo simulation can be used to find the resulting distribution of uncertainty. In the case where the uncertain variables are represented by independent stochastic variables given by expected value and standard deviation a linear approximation is used to calculate the resulting expected value and standard deviation.

An alternative approach is offered by **possibility** theory [2] based on representation of uncertain parameters by fuzzy numbers [3], [4]. This approach is well suited to represent lack of knowledge or imprecision in connection with economic parameters. The simplest fuzzy number being the interval, calculating with intervals and interval functions is far from trivial: In the case of an interval [5], [6] function being non-monotonic or variables appearing more than once, algorithms for finding global extreme points must be applied [7]-[9]. Since the basic operations when calculating with fuzzy numbers are interval operation the ability to perform correct calculations with intervals is a prerequisite for handling calculations with fuzzy numbers correctly. The interval approach

was previously applied to a product development case [10].

In this paper the two alternative approaches are compared based on a number of practical examples by means of numerically identical representations of input variables. For example, a triangular input parameter is interpreted as a triangular probability distribution when applying the probabilistic approach and a triangular fuzzy number when applying the possibility approach. As a general result, it is observed that the probabilistic approach results in numerically much smaller uncertainties than does the possibilistic approach. In view of the mega-project experience of budgets overruns, a critical discussion of weaknesses of conventional applied methods is presented and special attention is paid to the handling of outcomes with low probability/possibility but heavy impact.

It is further proposed to represent economic uncertainty by means of **imprecise stochastic variables** (or fuzzy random variables), thereby combining the probabilistic and possibilistic approaches in a unified concept. A railway reconstruction case is used to demonstrate the potential of imprecise stochastic variables offering a wide range of interpretations of the uncertainties represented.

2 Representation of uncertainty

For the comparative purpose of this paper, we introduce the rectangular, triangular, and trapezoidal representation of uncertain parameters. In each case we interpret the representation as a fuzzy number and a probability distribution as well

2.1 Rectangular representation

An uncertain variable X is represented by two real numbers a and b, where $a < b$. Interpreting X as a rectangular fuzzy number (actually an interval) we write

$$X = [a; b] \tag{1}$$

with the membership function

$$\begin{aligned} f(x) &= 1, \quad a \le x \le b, \\ &= 0, \quad otherwise. \end{aligned} \tag{2}$$

Interpreting X as a rectangular probability distribution with probability density function $f(x)$ (except for normalisation) we define the stochastic variable

$$X = \{\mu; \sigma\} \tag{3}$$

where μ is the expected or mean value and σ is the standard deviation given by

$$\mu = (a + b)/2, \quad \sigma^2 = (b - a)^2/12. \tag{4}$$

By normalisation the probability is constant in the interval $[a; b]$, equal to

$$h = 1/(b - a), \tag{5}$$

and equal to zero outside of the interval $[a; b]$.

2.2 Triangular representation

Here we represent an uncertain variable X by three real numbers a, c, and b, where $a < c < b$. Interpreting X as a triangular fuzzy number [11] we write

$$X = [a; c; b] \tag{6}$$

with the membership function

$$\begin{aligned} f(x) &= (x-a)/(c-a), && a \le x \le c, \\ &= (b-x)/(b-c), && c \le x \le b, \\ &= 0, && otherwise. \end{aligned} \tag{7}$$

X may also be interpreted as a triangular probability distribution function with probability density function (7) (except for normalisation) and we get for the mean value and standard deviation of the stochastic variable of the form (3)

$$\mu = (a+b+c)/3, \quad \sigma^2 = (a^2+b^2+c^2-ab-ac-bc)/18. \tag{8}$$

By normalisation the maximum probability h is attained at c,

$$h = 2/(b-a), \tag{9}$$

Outside of $[a; b]$ the probability is zero.

2.3 Trapezoidal representation

An uncertain variable X is represented by four real numbers a, c, d, and b, where $a < c < d < b$. This variable can be interpreted as a trapezoidal fuzzy number [12]. We write

$$X = [a; c; d; b] \tag{10}$$

and the membership function $f(x)$ is

$$\begin{aligned} f(x) &= (x-a)/(c-a), && a \le x \le c, \\ &= 1, && c \le x \le d, \\ &= (b-x)/(b-d), && d \le x \le b, \\ &= 0, && otherwise. \end{aligned} \tag{11}$$

X may also be interpreted as a trapezoidal probability distribution [13] with probability density function (11) (except for normalisation) and we get for the mean value and standard deviation of the stochastic variable of the form (3)

$$\begin{aligned} \mu &= h((b^3-d^3)/(b-d) - (c^3-a^3)/(c-a))/6 \\ \sigma^2 &= (3(r+2s+t)^4 + 6(r^2+t^2)(r+2s+t)^2 - \\ &\qquad (r^2-t^2)^2)/(12(r+2s+t))^2, \\ r &= c-a,\ s = d-c,\ t = b-d. \end{aligned} \tag{12}$$

In (11), the maximum probability h by normalisation is constant in the interval $[c; d]$,

$$h = 2/(b-a+d-c). \tag{13}$$

The probability is equal to zero outside of the interval $[a; b]$.

3 Processing of uncertain input variables

The actual economic problem under consideration is modelled by a function Y of n uncertain input variables

$$Y = Y(X_1,\ X_2, \ldots,\ X_n) \tag{14}$$

where Y is the output variable. With intervals and fuzzy numbers as input variables, the output variable is also an interval or a fuzzy number. When the input variables are probability distributions or stochastic variables so is the output variable.

3.1 Processing of intervals and fuzzy numbers

For basic operations on the intervals $X_1 = [a_1; b_1]$ and $X_2 = [a_2; b_2]$ we get the resulting interval Y by the formulas

$$\begin{aligned}
&Y = X_1 + X_2 = [a_1 + a_2; b_1 + b_2],\\
&Y = X_1 - X_2 = [a_1 - b_2; b_1 - a_2],\\
&Y = X_1 \cdot X_2\\
&\quad = [\min(a_1 \cdot a_2, a_1 \cdot b_2, b_1 \cdot a_2, b_1 \cdot b_2); \max(a_1 \cdot a_2, a_1 \cdot b_2, b_1 \cdot a_2, b_1 \cdot b_2)],\\
&Y = X_1/X_2\\
&\quad = [\min(a_1/a_2, a_1/b_2, b_1/a_2, b_1/b_2); \max(a_1/a_2, a_1/b_2, b_1/a_2, b_1/b_2)],\\
&\qquad\qquad 0 \notin [a_2; b_2].
\end{aligned} \tag{15}$$

It can be shown that the four basic interval operations are inclusion monotonic, commutative, and associative. However, the distributive rule is not valid in general. Instead, the so-called sub-distributivity holds, but only for addition and multiplication [7].

From a rational real valued function y of n real valued variables

$$y = y(x_1,\ x_2, \ldots,\ x_n) \tag{16}$$

we can create the interval extension function as an interval function Y of n intervals

$$Y = Y(X_1,\ X_2, \ldots,\ X_n) \tag{17}$$

simply by replacing the real operators by interval operators and the real variables by intervals.

A rational function can be formulated in many ways whereas the same reformulations cannot be done for interval expressions due to the invalidity of the distributive rule. This implies that different formulations of a rational function will lead to different interval extension functions and thus to different interval results [7]. In the case of Y being a monotonic function within the entire range of the input variables the minimum and maximum of Y as an interval can simply be found among the function values y at the extreme points of the variables. In the general case of Y being non-monotonic, variables appearing more then once, or intermediate variables are used, the calculation of Y as an interval is non-trivial.

In order to calculate correct results of interval extension functions in the general case we thus have to use global optimization methods. In this paper we use the program Interval Solver 2000 as an add-in module to MS-Excel for all interval calculations [14], [15]. Straightforward application of interval arithmetic (15) will result in excessive width output intervals.

Similar to the interval arithmetic formulas (15), we have for basic operations on triple estimate triangular fuzzy numbers $X_1 = [a_1; c_1; b_1]$ and $X_2 = [a_2; c_2; b_2]$ we get the resulting interval Y by the formulas (similar formulas exist for quadruple estimate trapezoidal numbers [16]).

$$\begin{aligned}
&Y = X_1 + X_2 = [a_1 + a_2; c_1 + c_2; b_1 + b_2],\\
&Y = X_1 - X_2 = [a_1 - b_2; c_1 - c_2; b_1 - a_2],\\
&Y = X_1 \cdot X_2\\
&\quad = [\min(a_1 \cdot a_2, a_1 \cdot b_2, b_1 \cdot a_2, b_1 \cdot b_2); c_1 \cdot c_2;\\
&\qquad \max(a_1 \cdot a_2, a_1 \cdot b_2, b_1 \cdot a_2, b_1 \cdot b_2)],\\
&Y = X_1 / X_2\\
&\quad = [\min(a_1/a_2, a_1/b_2, b_1/a_2, b_1/b_2); c_1/c_2;\\
&\qquad \max(a_1/a_2, a_1/b_2, b_1/a_2, b_1/b_2)],\ 0 \notin [a_2; b_2].
\end{aligned} \tag{18}$$

Mathematical operations on triangular fuzzy numbers can be facilitated by introducing the left $L(\alpha)$ and right $R(\alpha)$ representation. For a triangular fuzzy number with piece wise linear membership function we get

$$Y = [L(\alpha); R(\alpha)],\ \text{where } L(\alpha) = a + (c - a)\alpha \text{ and } R(\alpha) = b + (c - b)\alpha,\ \alpha \in [0, 1]. \tag{19}$$

For a trapezoidal fuzzy number we have correspondingly

$$Y = [L(\alpha); R(\alpha)],\ \text{where } L(\alpha) = a + (c - a)\alpha \text{ and } R(\alpha) = b + (d - b)\alpha,\ \alpha \in [0, 1]. \tag{20}$$

Observe that in this notation a fuzzy number is written as an interval with upper and lower bounds depending on α. This means that addition, subtraction, multiplication, and division can be carried out by using interval methods for all values of α. Likewise, for any triangular and trapezoidal function, the resulting triangular and trapezoidal functional values can be calculated and represented by L and R functions using interval methods for all values of α.

With triangular and trapezoidal fuzzy numbers as input variables the resulting membership function of the output variable Y is obtained by interval calculations on the α-cuts for a sufficient number of values of α, $0 \leq \alpha \leq 1$.

3.2 Processing of stochastic variables and probability distributions

With independent stochastic input variables Y is approximated by means of a Taylor series

$$\begin{aligned}
Y \cong Y(\mu_1, \ldots, \mu_n) + \partial Y/\partial X_1 \cdot (X_1 - \mu_1) + \partial Y/\partial X_2 \cdot (X_2 - \mu_2) + \ldots\\
+ \partial Y/\partial X_n \cdot (X_n - \mu_n)
\end{aligned} \tag{21}$$

where $\partial Y/\partial X_i$ is the partial derivative of Y with respect to X_i calculated at $(\mu_1, \ldots, \mu_n)$. The expected value is

$$E(Y) = \mu = Y(\mu_1, \ldots, \mu_n). \tag{22}$$

The standard deviation σ is approximated by

$$\sigma^2 \cong (\partial Y/\partial X_1)^2 \cdot \sigma_1^2 + \ldots + (\partial Y/\partial X_n)^2 \cdot \sigma_n^2. \tag{23}$$

In three cases, though, it is recommended to use Monte Carlo simulation:

1. When the uncertain variables are not statistically independent, use Monte Carlo simulation with a suitable covariance matrix.
2. When the function (16) is not monotonic and the linear approximation (21) therefore may be too inaccurate.
3. When the uncertain input variables are represented by probability distributions and the probability distribution of the output variable is needed.

4 Comparison of possibilities and probabilities

4.1 Non-monotonic test function

The non-monotonic real valued function

$$y = x(1 - x) \tag{24}$$

is calculated with rectangular argument $a = 0$ and $b = 1$.

Substituting x with a fuzzy number $X = [0; 1]$ we get by straightforward application of interval arithmetic (15) $Y = [0; 1]$. However, using global optimisation we get the result $Y = [0; 0, 25]$, which is the correct result, i.e. the most narrow interval that can be obtained for Y.

Interpreting x as a uniform probability distribution we get by Monte Carlo simulation [17] the results show in Fig. 1, where also the membership function of the above result is depicted (normalised as a pdf). It is seen that the Monte Carlo simulation reproduces the minimum and maximum limitations of the function. However, the shapes of the membership function and the probability distribution are quite different.

4.2 Sum of uncertain variables

Project cost functions are often sums of uncertain variables. As an example consider a simple sum function of 10 uncertain cost variables represented by identical triangular representations, $a = 8$, $c = 10$, and $b = 16$. Computational results are shown in Fig. 2. First, observe the triangular piecewise linear membership obtained by fuzzy arithmetic (15). First, compare with the two Monte Carlo simulations, one with uncorrelated variables, the other with 100% correlation. The former is well approximated with a normal distribution. The latter

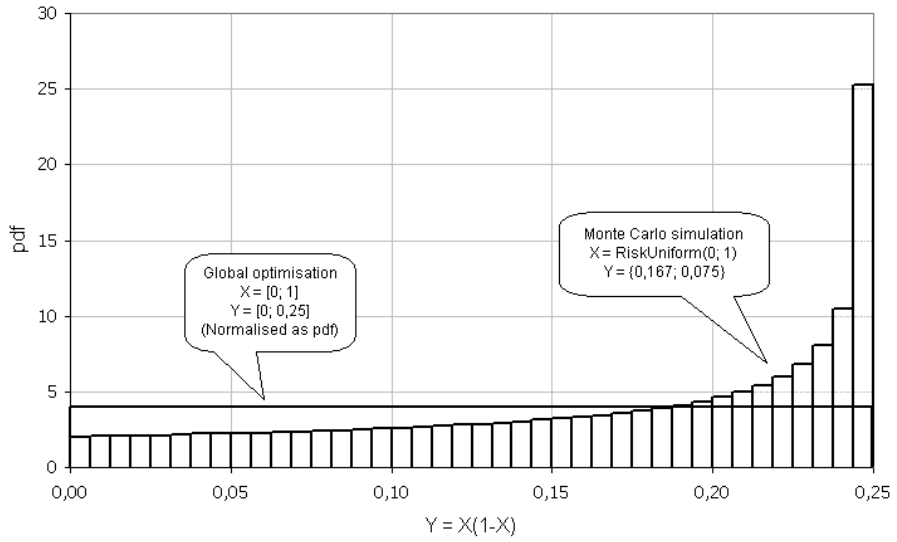


Figure 1: Function x(1-x) with rectangular representation of x: a = 0, b = 1.

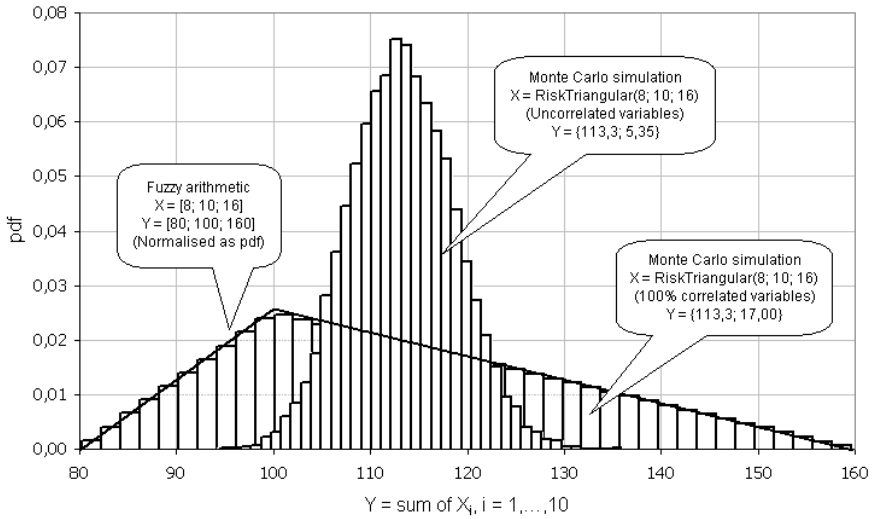


Figure 2: Sum of 10 identical cost elements with triangular representation of x: a = 8, c = 10, b = 16.

exactly reproduces the triangular shape of the fuzzy membership function.

From a cost uncertainty point of view the approaches give rise to alternative conclusions. The "probabilist" argues that the total cost of the project most probably will be 113,3, provided the variables are independent. The probability of overrunning the expected total cost with a certain amount is equal to the probability of running lower. The "possibilist" argues that the total cost of the project is 100 and the possibility of overrunning is considerably larger than the

opposite. Also note that the total cost expected by the "probabilist" is $13,3\%$ higher that the one expected by the "possibilist". It is interesting to note that the "probabilist" will have to accept that in practice the variables are not uncorrelated, although he might have difficulties in determining the correlation coefficients. Actually, assuming 100% correlation he discovers that he is facing exactly the same numerical uncertainty as the "possibilist" because the probability distribution is coinciding with the membership function. It should be noted here that this coincidence is also happening when calculating differences. For multiplication and division things are more complex.

Simularly, using trapezoidal cost elements, $a = 8$, $c = 9$, $d = 11$, and $b = 16$, we get the fuzzy total cost $Y = [80; 90; 110; 160]$ and the probability total cost $Y = \{112;\ 5,58\}$ normally distributed by independent input variables.

5 Imprecise stochastic variables

The representation of uncertain variables by a conventional stochastic approach is generally accepted to account for uncertainties of a statistical nature. In case the expected value of an uncertain economic parameter represented by a conventional stochastic variable (3) is known only with imprecision, we propose to represent it by an imprecise stochastic variable $\mathbf{X}$

$$\mathbf{X} = \{\boldsymbol{\mu}; \sigma\}, \tag{25}$$

where the expected value $\boldsymbol{\mu}$ is now a fuzzy number accounting for the **imprecision** of the actual economic parameter. The **variability** is still precisely accounted for by the standard deviation σ.

In (25) the fuzzy expected value $\boldsymbol{\mu}$ may have different forms, e.g. an interval

$$\boldsymbol{\mu}^I = [a; b], \tag{26}$$

or a triple estimate corresponding to a triangular fuzzy number with α-cuts 0 and 1

$$\boldsymbol{\mu}^T = [a; c; b], \tag{27}$$

or even a quadruple estimate corresponding to a trapezoidal fuzzy number with α-cuts 0 and 1

$$\boldsymbol{\mu}^{TR} = [a; c; d; b]. \tag{28}$$

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be independent stochastic variables with expected values $E(\mathbf{X}_1) = \boldsymbol{\mu}_1$ and $E(\mathbf{X}_2) = \boldsymbol{\mu}_2$ and variances $\mathrm{VAR}(\mathbf{X}_1) = \sigma_1^2$ and $\mathrm{VAR}(\mathbf{X}_2) = \sigma_2^2$. We then have for the basic calculations with $\mathbf{X}_1$ and $\mathbf{X}_2$:

$$\begin{array}{ll} \textit{Addition}: & \boldsymbol{\mu} = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2; \sigma^2 = \sigma_1^2 + \sigma_2^2. \\ \textit{Subtraction}: & \boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2; \sigma^2 = \sigma_1^2 + \sigma_2^2. \\ \textit{Multiplication}: & \boldsymbol{\mu} = \boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_2; \boldsymbol{\sigma}^2 \cong \sigma_1^2 \cdot \boldsymbol{\mu}_2^2 + \sigma_2^2 \cdot \boldsymbol{\mu}_1^2. \\ \textit{Division}: & \boldsymbol{\mu} = \boldsymbol{\mu}_1 / \boldsymbol{\mu}_2; \boldsymbol{\sigma}^2 \cong \sigma_1^2 / \boldsymbol{\mu}_2^2 + \sigma_2^2 \cdot \boldsymbol{\mu}_1^2 / \boldsymbol{\mu}_2^4, \text{ if } 0 \notin \boldsymbol{\mu}_2. \end{array} \tag{29}$$

6 A practical cost estimation case

Consider the case of estimating the total cost incurred by a railway reconstruction project described by independent imprecise stochastic variables, namely 18 cost items $\mathbf{X}_1, \ldots, \mathbf{X}_{18}$ and 3 correction factors $\mathbf{X}_{19}, \ldots, \mathbf{X}_{21}$. The correction factors are introduced in order to account for overall influences not accounted for by the individual cost items.

The total cost before corrections is the sum

$$\mathbf{Y}_1 = \mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_{18}. \tag{30}$$

The total cost after corrections $\mathbf{Y} = \mathbf{Y}(\mathbf{X})$ is a non-linear function of all 21 stochastic variables

$$\mathbf{Y} = (\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_{18}) \cdot \mathbf{X}_{19} \cdot \mathbf{X}_{20} \cdot \mathbf{X}_{21}. \tag{31}$$

The variability and imprecision of the case represented by standard deviations and expected values of the 21 input parameters are estimated by railway experts with relevant project experience. Subsequently, the expected value and standard deviation of the total cost $\mathbf{Y}$ is calculated by means of extended application of (4) and (5) respectively. (Similar, yet simpler, formulas hold for Y_1):

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 + \ldots + \boldsymbol{\mu}_{18}) \cdot \boldsymbol{\mu}_{19} \cdot \boldsymbol{\mu}_{20} \cdot \boldsymbol{\mu}_{21} \tag{32}$$

and

$$\boldsymbol{\sigma} \cong (\partial \mathbf{Y}/\partial \mathbf{X}_1)^2 \cdot \boldsymbol{\sigma}_1^2 + \ldots + (\partial \mathbf{Y}/\partial \mathbf{X}_n)^2 \cdot \boldsymbol{\sigma}_n^2 \tag{33}$$

where

$$\partial \mathbf{Y}/\partial \mathbf{X}_i = \mathbf{X}_{19} \cdot \mathbf{X}_{20} \cdot \mathbf{X}_{21}, \qquad i = 1, \ldots, 18, \tag{34}$$

and

$$\begin{aligned} \partial \mathbf{Y}/\partial \mathbf{X}_{19} &= (\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_{18}) \cdot \mathbf{X}_{20} \cdot \mathbf{X}_{21} \\ \partial \mathbf{Y}/\partial \mathbf{X}_{20} &= (\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_{18}) \cdot \mathbf{X}_{19} \cdot \mathbf{X}_{21} \\ \partial \mathbf{Y}/\partial \mathbf{X}_{21} &= (\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_{18}) \cdot \mathbf{X}_{19} \cdot \mathbf{X}_{20} \end{aligned} \tag{35}$$

The total cost estimation results are shown in Table 1 together with a technical explanation of all the variables in terms of cost items and correction factors. The term "code" refers to the cost structure hierarchy. Focusing on total cost after corrections, an interpretation and discussion of the results with gradually increased uncertainty follows.

Imprecision and variation may be combined in the cost estimation by presenting the total cost after corrections according to Table 1 as e.g.

$$\mathbf{Y} = \{[23.842;\ 32.930];\ 3.659\} \tag{36}$$

or

$$\mathbf{Y} = \{[23.842;\ 26.565;\ 32.930];\ 3.659\} \tag{37}$$

Var.	Code	Item	$\{\mu; \sigma\}$
	0, 00	**Management and specs.**	{[1.732; 1.780; 1.884]; 268}
$\mathbf{X}_1$	0.10	Project management	{[524; 540; 575]; 160}
$\mathbf{X}_2$	0.20	Construction management etc.	{[975; 1.000; 1.050]; 200}
$\mathbf{X}_3$	0.30	Design specifications etc.	{233; 240; 259]; 80}
$\mathbf{X}_4$	**10, 00**	**Environmental and soil eng.**	{[864; 888; 950]; 194}
$\mathbf{X}_5$	**20, 00**	**Traffic tasks**	{[48; 50; 53]; 12}
	30, 00	**Renewal of tracks**	{[8.907; 9.190; 9.664]; 383}
$\mathbf{X}_6$	30,10	New outbound main track	{[975; 1.000; 1.050]; 200}
$\mathbf{X}_7$	30,20	Track renewal at platform 3/5	{[5.432; 5.600; 5.880]; 300}
$\mathbf{X}_8$	30,30	New platform edge	{[1.533; 1.580; 1.643]; 50}
$\mathbf{X}_9$	30,40	Track renewal depot, West	{[285; 300; 321]; 80}
$\mathbf{X}_{10}$	30,50	Track layout design	{[682; 710; 770]; 90}
$\mathbf{X}_{11}$	**40, 00**	**Platform and station**	{[538; 560; 602]; 120}
$\mathbf{X}_{12}$	**50, 00**	**Safety and signal installations**	{[5.035; 5.245; 5.586]; 1.428}
	60, 00	**Informatics incl. power supply**	{[2.374; 2.417; 2.626]; 221}
$\mathbf{X}_{13}$	60,10	Phase 2-4	{[78; 80; 86]; 22}
$\mathbf{X}_{14}$	60,20	Sub project management	{[249; 259; 275]; 76}
$\mathbf{X}_{15}$	60,30	Passenger information	{[1.009; 1.030; 1.123]; 140}
$\mathbf{X}_{16}$	60,40	Electrical power supply	{[1.038; 1.048; 1.142]; 152}
	70, 00	**Overhead line incl. pylons**	{[3.507; 3.624; 3.787]; 487}
$\mathbf{X}_{17}$	70,10	Overhead cables	{[3.021; 3.122; 3.262]; 480}
$\mathbf{X}_{18}$	70,20	Layout and planning	{[486; 502; 525]; 82}
$\mathbf{Y}_1$		**Total cost before corrections**	{[23.005; 23.754; 25.152]; 1.611}
$\mathbf{X}_{19}$	A	Internal decision process	{[1, 006; 1, 032; 1, 098]; 0, 068}
$\mathbf{X}_{20}$	B	Design specifications etc.	{[1, 009; 1, 040; 1, 100]; 0, 068}
$\mathbf{X}_{21}$	C	Working process	{[1, 021; 1, 042; 1, 084]; 0, 079}
$\mathbf{Y}$		**Total cost after corrections**	{[23.842; 26.565; 32.930]; 3.659}

Table 1: Total cost estimation for railway reconstruction case by imprecise stochastic variables (1000 monetary units)

Some comments concerning the actual shape of the cumulated distribution function (cdf) connected with the imprecise stochastic variables of the total cost before and after corrections are in order. Since the cost function (31) basically is a sum of many small independent contributions (of unspecified shapes) it follows from the central limit theorem that the resulting distribution may be well approximated by a normal distribution.

By closer inspection of the set of conventional normal distributions generated by (37) a number of observations concerning the uncertainty of the total cost after corrections can be made:

1. Ignoring both variability and imprecision, the conventional (crisp) value of the total cost after corrections is 26.565, which represents a conventional budget without taking uncertainties into account.

2. Ignoring variability the double estimate of the total cost after corrections is [23.842; 32.930]. This accounts for the imprecision involved in the cost calculation.

3. Ignoring variability the triple estimate of the total cost after corrections is [23.842; 26.565; 32.930]. In this way the conventional budget figure of the total cost is represented in the uncertainty estimate.

4. Next consider the cost function represented by the normal distribution $\{26.565; 3.659\}$, thus ignoring imprecision. By inspection of the cumulative distribution function the following statement can be made: With probability 0,9 the total cost after corrections will be below ~ 31.000.

5. Next consider the normal distribution functions $\{[23.842; 32.930]; 3.659\}$, thus taking imprecision of the expected values into account. This means that in the worst case the total cost after corrections will be below ~ 37.500 with probability 0,9 and in the best case below ~ 28.500 with probability 0,9.

7 Conclusion

The results presented in this paper indicate that the picture of economic uncertainty very much depends on the way uncertainty is represented and processed. The point of departure is the availability of a conventional economic model, e.g. in the form of a project budget. The two alternative ways of representing uncertainty, namely possibilities and probabilities, can be considered as extensions of a conventional budget. In the case of possibilities being modelled by fuzzy numbers, correct calculations require application of global optimisation. By using straightforward fuzzy arithmetic, the resulting uncertainty is incorrectly getting excessively large. In the case of uncertainty being modelled by stochastic variables or probability distributions, linear approximation or Monte Carlo simulation is applicable. Specific attention has to be paid to correlation between probabilistic variables: Uncertainty tend to become much larger with correlated variables compared to independent variables. Based on the calculations done for a sum of 10 triangular uncertain variables it is demonstrated that by assuming 100% correlation the total uncertainty exactly reproduces the

result obtained by fuzzy calculations.

Fundamental for the comparison of probabilistic and possibilistic approaches in this paper is the usage of numerically identical uncertain input variables, namely rectangular, triangular, and trapezoidal uncertainty. In case of skewed input variables, it is observed that the conventional budget values are not preserved when probabilities are applied, contrary to the case of fuzzy numbers. Further, it seems intuitively strange that the resulting probabilistic uncertainty is symmetric (a normal distribution), considering the fact, that more often budgets are overrun than the opposite.

The introduction of imprecise stochastic variables allows for simultaneous representation of imprecision and variability. As demonstrated by a railway reconstruction project imprecise stochastic variables allow for a wide range of uncertainty representations and calculations. Basically, uncertainty of a statistical nature as well as uncertainty of a non-statistical nature can be represented, calculated, and communicated by means of a unified concept.

Modelling of economic uncertainty is crucially dependent on reliable input data. This aspect, however, has not been dealt with in the present paper but is the subject of an ongoing research project. An initiative taken by the Danish Government instigates a new way of dealing with risk and uncertainty in large infrastructure projects that is primarily based of objective, rather than subjective uncertainty data. It is expected that the concepts outlined in this paper will find areas of application in that particular context.

References

[1] Lichtenberg, S. (2000), *Proactive management of uncertainty using the successive principle.* Polyteknisk Press, Copenhagen.

[2] Dubois, D. and Prade, H. (1998), *Possibilty theory - An approach to computerized processing of uncertainty.* New York: Plenum Press.

[3] Dubois, D. and Prade, H. (1978), Operations on fuzzy numbers. *International Journal of System Science*, 9: 613-626.

[4] Dubois, D. and Prade, H. (1979), Fuzzy real algebra: Some results. *Fuzzy Sets and Systems*, 2: 327-348.

[5] Moore, R.E. (1962), *Interval arithmetic and automatic error analysis in digital computing.* Ph.D. dissertation, Stanford University, USA.

[6] Moore, R.E. (1966), *Interval analysis.* New Jersey: Prentice-Hall.

[7] Caprani, O., Madsen, K. and Nielsen H.B. (2002), *Introduction to interval analysis.* Lecture notes, Department of Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

[8] Hansen, E. (1992), *Global optimization using interval analysis.* Marcel Dekker, New York, USA.

[9] Kjřller, S., Kozine, P., Madsen, K. and Stauning O. (2007), *Non-linear global optimization using interval arithmetic and constraint propagation.* In: A. Törn and J. Zilinskas (Eds.), Models and algorithms for global optimization, Springer Verlag.

[10] SchjÊr-Jacobsen, H. (1996), A new method for evaluating worst- and best-case economic consequences of technological development. *International Journal of Production Economics*, 46-47: 241-250.

[11] Chiu, C.U. and Park, C.S. (1994), Fuzzy cash flow analysis using present worth criterion. *The Engineering Economist*, 39(2): 113-138.

[12] Wang, M.-J. and Liang, G.-S. (1995), Benefit/cost analysis using fuzzy concept. *The Engineering Economist,* 40(4), 359-376.

[13] Kacker, R. and Lawrence, J. (2007), Trapezoidal, triangular, and rectangular distributions for Type B evaluations, *Metrologia,* 44(2): 117-127.

[14] Hyvönen, E. and de Pascale, S. (1999), A new basis for spreadsheet computing: Interval Solver for Microsoft Excel. *Proceedings of the 11th Innovative Applications of Artificial Intelligence (IAAI-99)*, AAAI Press, Menlo Park, California.

[15] Hyvönen, E. and de Pascale, S. (2000), *Interval Solver 2000 for Microsoft Excel.* User's Guide, Version 4.0. Program Release 4.0.0.2, Delisoft Ltd., Helsinki, Finland.

[16] SchjÊr-Jacobsen, H. (2004), Modeling of economic uncertainty, *Fuzzy Economic Review*, IX(2): 49-73.

[17] Palisade Corporation, (2008), *@Risk 5.0, Monte Carlo simulation add-in module for MS Excel,* www.palisade.com/risk/.

On Open Questions in the Geometric Approach to Learning BN Structures

Milan Studený
Dept. of Decision-Making Theory
ÚTIA AV ČR, v.v.i.
studeny@utia.cas.cz

Jiří Vomlel
Dept. of Decision-Making Theory
ÚTIA AV ČR, v.v.i.
vomlel@utia.cas.cz

Abstract

The basic idea of an algebraic approach to learning Bayesian network (BN) structures is to represent every BN structure by a certain (uniquely determined) vector, called the *standard imset*. In a recent paper [11], we have shown that the set S of standard imsets is the set of vertices (= extreme points) of a certain polytope P and introduced natural *geometric neighborhood* for standard imsets, and, consequently, for BN structures.

The new geometric view led to a series of open mathematical questions. In this contribution, we try to answer some of them. First, we introduce a class of necessary linear constraints on standard imsets and formulate a conjecture that these constraints characterize the polytope P. The conjecture has been confirmed in the case of (at most) 4 variables. Second, we confirm a former hypothesis by Raymond Hemmecke that the only lattice points (= vectors having integers as components) within P are standard imsets. Third, we give a partial analysis of the geometric neighborhood in the case of 4 variables.

1 Motivation

The motivation for this research is learning Bayesian network (BN) structures from data by the method of maximization of a quality criterion (= the score and search method). By a *quality criterion* is meant a real function $\mathcal{Q}$ of the BN structure (= of a graph G, usually) and of the database D. The value $\mathcal{Q}(G, D)$ should say how much the BN structure given by G is good to explain the occurrence of the database D.

The basic idea of an algebraic and geometric approach to this topic, proposed in Chapter 8 of [8] and then developed in [11], is to represent the BN structure given by an acyclic directed graph G by a certain vector u_G having integers as components, called the *standard imset* (for G). The point is that then every reasonable criterion $\mathcal{Q}$ for learning BN structures (score equivalent and decomposable one) is an affine function (= a linear function plus a constant) of the standard imset. More specifically, one has

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

where $s_D^{\mathcal{Q}}$ is a real number, $t_D^{\mathcal{Q}}$ a vector of the same dimension as the standard imset u_G (these parameters both depend solely on the database D and the

criterion $\mathcal{Q}$) and $\langle *, * \rangle$ denotes the scalar product. The vector $t_D^{\mathcal{Q}}$ is named the *data vector* (relative to $\mathcal{Q}$).

The main result of [11] 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. Thus, as every reasonable quality criterion $\mathcal{Q}$ can be viewed as (the restriction of) an affine function on the respective Euclidean space (of higher dimension), the task to maximize $\mathcal{Q}$ over BN structures is equivalent to the task to maximize an affine function over the above-mentioned polytope P.

This maximization problem has been treated thoroughly within the linear programming community. A classic tool to solve linear programming problems is the *simplex method* [5]. One of possible interpretations of this method is that it is a kind of a search method, in which one moves between the vertices of the polytope along its edges (in the geometric sense) until an optimal vertex is reached. This motivated the concept of the *geometric neighborhood* for standard imsets, and, consequently, for BN structures.

Several open mathematical questions have been mentioned in the conclusions of [11]. They are motivated by the above-mentioned intention to apply linear programming methods in the area of learning BN structures. This contribution is devoted to three of them.

2 Basic concepts

2.1 Learning BN structures

Throughout this paper we assume that N is a non-empty finite set of *variables*. Every variable $i \in N$ is assigned a finite set of possible values, the individual sample space X_i. To avoid trivial cases and consequent troubles we assume $|\mathsf{X}_i| \geq 2$ for any $i \in N$.

Let $\mathsf{DAGS}(N)$ denote the collection of all acyclic directed graphs having N as the set of nodes. The (discrete) *Bayesian network* (BN) is a pair (G, P), where $G \in \mathsf{DAGS}(N)$ and P is a probability distribution on the joint sample space $\mathsf{X}_N \equiv \prod_{i \in N} \mathsf{X}_i$ which (recursively) factorizes according to G [4]. Given $G \in \mathsf{DAGS}(N)$, the respective statistical model of a *BN structure* is the class of all distributions P on X_N that factorize according to G.

Note it may happen that two different graphs over N describe the same BN structure. Thus, one is usually interested in describing the BN structure by a unique representative. A classic such graphical representative is a special chain graph, called the *essential graph* [1]. However, in our algebraic approach, we use an algebraic representative instead, called the *standard imset* (see below). There is a polynomial algorithm for transforming the essential graph into the standard imset and conversely [10].

Learning BN structures is done on the basis of data, assumed in the form of a complete database $D : x^1, \ldots, x^d$ of the length $d \geq 1$, which is a sequence of elements of the joint sample space X_N. Let $\mathsf{DATA}(N, d)$ denote the collection of all databases from X_N of the length d. A *quality criterion* (for learning BN structures) is a real function $\mathcal{Q}$ on $\mathsf{DAGS}(N) \times \mathsf{DATA}(N, d)$. The learning procedure based on $\mathcal{Q}$ consists in maximizing the function $G \mapsto \mathcal{Q}(G, D)$ over $G \in \mathsf{DAGS}(N)$, where $D \in \mathsf{DATA}(N, d)$ is the observed database. Thus, the value $\mathcal{Q}(G, D)$ should somehow evaluate how the statistical model determined

by G fits the database D. We refer for the related concept of (statistical) consistency of a quality criterion to § 8.4.2 in [4].

However, there are other technical requirements on quality criteria raised in connection with computational methods for their maximization [3]. A criterion is *decomposable* if it is the sum of contributions that correspond to factors in the factorization according to the graph and *score equivalent* [2] if it ascribes the same value to graphs describing the same BN structure. There are several examples of quality criteria that meet these requirements. A kind of standard example of such a criterion is Schwarz's *Bayesian information criterion* (BIC) [6], but there is also a bunch of Bayesian quality criteria [9].

2.2 A few concepts from polyhedral geometry

Let us consider a real Euclidean space $\mathbb{R}^K$, where K is a non-empty finite set. The scalar product of two vectors $\boldsymbol{v}, \boldsymbol{x}$ in $\mathbb{R}^K$ will be denoted as follows:

$$\langle \boldsymbol{v}, \boldsymbol{x} \rangle \equiv \sum_{s \in K} v_s \cdot x_s \,.$$

A *rational polytope* in $\mathbb{R}^K$ is the convex hull of a finite set $\mathsf{V} \subseteq \mathbb{Q}^K$ of rational points. A well-known result in polyhedral geometry (Corollary 7.1c in [5]) says that a polytope can equivalently be characterized by means of a finite number of linear inequality constraints.

Note that the classic version of the *simplex method* is applicable to the task to find maximum/minimum of a linear function over a set $\mathsf{P} \subseteq \mathbb{R}^K$ defined by means of a finite number of linear inequality constraints (see Chapter 11 in [5]).

A *vertex* (= an extreme point) of a polytope P is a point in P which cannot be written as a convex combination of other elements in P. An *edge* of P is a line-segment $[\boldsymbol{x}, \boldsymbol{y}]$, where $\boldsymbol{x}, \boldsymbol{y}$ are distinct vertices of the polytope P and the set $\mathsf{P} \setminus [\boldsymbol{u}, \boldsymbol{v}]$ is convex. The vertices and edges of a polytope are quite important in linear programming because the simplex method applied to a polytope P can be interpreted as a kind of search method in which one moves between the vertices of P along its (geometric) edges (see § 11.1 of [5]).

2.3 Imsets

The method of *structural imsets* has been proposed in [8] to provide an universal (mathematical) tool for describing probabilistic conditional independence structures. In the context of graphical models, it leads to an algebraic approach to learning BN structures.

An *imset* u over N is an integer-valued function on $\mathcal{P}(N) \equiv \{A;\, A \subseteq N\}$, the power set of N. It can be viewed as a vector whose components are integers, indexed by subsets of N. Any real function $m : \mathcal{P}(N) \to \mathbb{R}$ will be analogously interpreted as a real vector (= identified with an element of $\mathbb{R}^{\mathcal{P}(N)}$). Thus, an imset is nothing but an element of $\mathbb{Z}^{\mathcal{P}(N)}$; in the context of integer programming [5] called a *lattice point* in the Euclidean space $\mathbb{R}^{\mathcal{P}(N)}$.

A trivial example of an imset is the *zero imset*, denoted by 0. Given $A \subseteq N$, the symbol δ_A will denote this basic imset:

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

Since $\{\delta_A;\ A \subseteq N\}$ is a linear basis of $\mathbb{R}^{\mathcal{P}(N)}$, any imset can be expressed as a combination of these basic imsets.

An *elementary imset* (over N) is an imset of the form

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

where $C \subseteq N$ and $a, b \in N \setminus C$ are distinct. In our algebraic approach [8] it encodes an elementary conditional independence statement $a \perp\!\!\!\perp b \,|\, C$. The class of all elementary imsets over N will be denoted by $\mathcal{E}(N)$; it is a finite subset of $\mathbb{R}^{\mathcal{P}(N)}$. The cone spanned by $\mathcal{E}(N)$ will be denoted by $\mathcal{R}(N)$.[1]

An imset will be called *combinatorial* if it is a combination of elementary imsets with non-negative integers as coefficients.[2] The *degree* of a combinatorial imset u, denoted by $deg\,(u)$, is the number

$$deg\,(u) = \langle m_*, u\rangle \equiv \sum_{S\subseteq N} m_*(S)\cdot u(S)\,, \tag{1}$$

where $m_*(S) = \frac{1}{2}\cdot|S|\cdot(|S|-1)$ for $S \subseteq N$. It is shown in Proposition 4.3 of [8] that $deg\,(u)$ is the sum of coefficients in the decomposition of u into elementary imsets; in particular, this sum only depends on u, not on a particular combination of elementary imsets yielding u.

2.4 Algebraic approach to learning BN structures

Given $G \in \mathsf{DAGS}(N)$, the *standard imset* for G is given by the formula:

$$u_G \;=\; \delta_N - \delta_\emptyset + \sum_{i\in N} \{\,\delta_{pa_G(i)} - \delta_{\{i\}\cup pa_G(i)}\,\}, \tag{2}$$

where $pa_G(i) = \{\, j \in N;\ j \to i \ \text{ in } G\,\}$ denotes the set of *parents* if i in G. Note that the terms in (2) can both sum up and cancel each other. Nevertheless, it follows from the definition that u_G has at most $2\cdot|N|$ non-zero values. Thus, the memory demand for representing standard imsets are polynomial in $|N|$.

An important observation is that, for $G, H \in \mathsf{DAGS}(N)$, one has $u_G = u_H$ iff they describe the same BN structure (Corollary 7.1 in [8]). In particular, the standard imset for $G \in \mathsf{DAGS}(N)$ is a unique representative of the corresponding BN structure. Note that every standard imset is combinatorial; actually, it is a sum of elementary imsets (see Lemma 2 in § 5). The degree of a standard imset u_G is $\binom{|N|}{2} - r$, where r is the number of arrows in G (see Lemma 7.1 in [8]).

Now, Lemmas 8.3 and 8.7 from [8] together say that every score equivalent and decomposable criterion $\mathcal{Q}$ must have the form:

$$\mathcal{Q}(G, D) \;=\; s^{\mathcal{Q}}_D - \langle t^{\mathcal{Q}}_D, u_G\rangle \quad \text{for } G \in \mathsf{DAGS}(N), D \in \mathsf{DATA}(N, d), d \geq 1 \tag{3}$$

where the constant $s^{\mathcal{Q}}_D \in \mathbb{R}$ and the vector $t^{\mathcal{Q}}_D : \mathcal{P}(N) \to \mathbb{R}$ do not depend on G. The formulas for the data vector $t^{\mathcal{Q}}_D$ relative to some basic quality criteria $\mathcal{Q}$ have been derived in [8, 9].

---

[1] It is a pointed rational polyhedral cone in $\mathbb{R}^{\mathcal{P}(N)}$.

[2] Equivalently, the sum of elementary imsets with allowed repetition of summands.

2.5 Geometric view on learning BN structures

Let us take a geometric view on the set of standard imsets over a fixed set of variables N, denoted by S:

$$\mathsf{S} \equiv \{\, u_G;\, G \in \mathsf{DAGS}(N) \,\} \subseteq \mathbb{R}^{\mathcal{P}(N)}\,.^3$$

Theorem 4 in [11] says that S is the set of vertices of a rational polytope $\mathsf{P} \subseteq \mathbb{R}^{\mathcal{P}(N)}$, whose dimension is $2^{|N|} - |N| - 1$. This polytope P will be called the *standard imset polytope* in the sequel. It follows from (3) that the task to maximize $\mathcal{Q}$ over $G \in \mathsf{DAGS}(N)$ is equivalent to the task to minimize the linear function $u \mapsto \langle t_D^{\mathcal{Q}}, u\rangle$ over P.

The idea of application of linear programming methods in the area of learning BN structures led to the concept of geometric neighborhood for BN structures. More specifically, two standard imsets $u, v \in \mathsf{S}$ will be called the *geometric neighbors* if the line-segment connecting them in $\mathbb{R}^{\mathcal{P}(N)}$ is an edge of the standard imset polytope P.

It has been shown in Theorem 5 of [11] that the well-known *inclusion neighborhood*, used widely in present computational methods for learning BN structures, like the GES algorithm [3], is strictly contained in the geometric one. Moreover, it follows from Corollary 8.4 in [8] that standard imsets $u, v \in \mathsf{S}$ correspond to inclusion neighbors iff their differential imset $w = u - v$ is either elementary one or a multiple of it by -1.

The importance of the concept of geometric neighborhood is based on the fact that, for any affine function $\mathcal{Q}$ on P, a local maximum of $\mathcal{Q}$ in $u \in \mathsf{S}$ with respect to the geometric neighborhood must be the global maximum of $\mathcal{Q}$ over P (Theorem 6 in [11]). In particular, this holds for any reasonable quality criterion $\mathcal{Q}$ for learning BN structures. The following research goals have been expressed in conclusions of [11]:

- Describe linear constraints on the elements P. A complete characterization of these constraints would provide a description of P suitable for the intended application of linear programming methods.
- An interesting related conjecture by Raymond Hemmecke is that the only lattice points within P are standard imsets.
- Describe *differential imsets* for geometric neighbors, that is, imsets of the form $u_G - u_H$, where $G, H \in \mathsf{DAGS}(N)$ are such that u_G and u_H are geometric neighbors.

These questions concern the complexity of a potential future linear programming procedure for maximization of a quality criterion $\mathcal{Q}$. In this paper we answer partially some of them.

3 Necessary linear constraints

In this section, we summarize all linear constraints on standard imsets we are aware of. Of course, they give necessary conditions on points in P.

[3] To avoid misunderstanding recall that distinct $G, H \in \mathsf{DAGS}(N)$ may give rise the same standard imset $u_G = u_H$ but S contains just one vector for any group of graphs defining the same BN structure.

3.1 Overview of the constraints

We classify our linear constraints into three groups, denoted (A), (B) and (C). First, basic results from [8] imply that every standard imset belongs to the cone $\mathcal{R}(N)$ generated by elementary imsets. This observation implies two kinds of necessary linear conditions on the elements of P: the equality constraints, denoted by (A), and the remaining inequality constraints, denoted by (B).

(A) Equality constraints

If $u \in \mathsf{S}$ then the following two conditions are valid:

$$(A.1) \qquad \sum_{S,\, S\subseteq N} u(S) = 0,$$

$$(A.2) \qquad \forall\, i \in N \qquad \sum_{S,\, i\in S\subseteq N} u(S) = 0.$$

This means that S, and, therefore, P as well, belongs to a linear subspace of $\mathbb{R}^{\mathcal{P}(N)}$ of the dimension $2^{|N|} - |N| - 1$.

(B) Non-specific inequality constraints

The inequality constraints for points in the cone $\mathcal{R}(N)$ are related to supermodular functions. A function $m : \mathcal{P}(N) \to \mathbb{R}$ is called *supermodular* iff

$$m(C \cup D) + m(C \cap D) \geq m(C) + m(D) \qquad \text{for every } C, D \subseteq N.$$

An equivalent definition is that $\langle m, v\rangle \geq 0$ for every elementary imset v over N. This observation gives a (formally infinite) set of inequality constraints for the points in $\mathcal{R}(N)$, and, therefore, for any standard imset u:

$$(B) \qquad \langle m, u\rangle \geq 0 \quad \text{for every supermodular function } m : \mathcal{P}(N) \to \mathbb{R}.$$

Nevertheless, the point is that this condition can equivalently be formulated in the form of a finite number of linear inequality constraints. First, without loss of generality one can assume that $m(S) = 0$ for $S \subseteq N$ with $|S| \leq 2$. Second, the class of these special supermodular functions is a pointed rational polyhedral cone and has, therefore, finitely many extreme rays.[4] Thus, the class normalized integral representatives of these extreme rays, denoted by $\mathcal{K}^{\diamond}_{\ell}(N)$ and called the ℓ-skeleton in [8], establishes a finite set of normalized inequality constraints:

$$\forall\, m \in \mathcal{K}^{\diamond}_{\ell}(N) \qquad \langle m, u\rangle \geq 0.$$

These (representatives of) extreme rays have been computed for $|N| \leq 5$ using linear programming packages [7]. It seems that the number of these extreme rays grows super-exponentially with $|N|$; their numbers are in Table 1.

It looks like none of these inequality constraints for P is derivable from the other constraints (including those mentioned below).

[4] See § 5.1.2 and Lemma 5.3 (pp. 90-93) in [8] for both these claims.

Table 1: Number of non-specific inequality constraints.

$\lvert N\rvert$	2	3	4	5
$\lvert\mathcal{K}^{\diamond}_{\ell}(N)\rvert$	1	5	37	117978

(C) Specific inequality constraints

The results of [10] led to a series of specific linear inequality constraints for standard imsets, that are not valid for all points in the cone $\mathcal{R}(N)$. These constraints are related to "ascending" classes of sets. We say that a class $\mathcal{A} \subseteq \mathcal{P}(N)$ of subsets of N is *closed under supersets* if

$$\forall\, S \in \mathcal{A} \quad \text{if } S \subseteq T \subseteq N \quad \text{then } T \in \mathcal{A}.$$

To avoid vacuous constraints and a trivial consequence of (A.1) we consider only non-empty classes of non-empty sets. This gives the following series of constraints:

$$(C) \qquad \sum_{S\in\mathcal{A}} u(S) \leq 1 \quad \text{for any system } \emptyset \neq \mathcal{A} \subseteq \{S \subseteq N; |S| \geq 1\} \text{ which is closed under supersets.}$$

Note that, unlike the number (B)-constraints, the number of constraints in (C) seems to grow only exponentially with $|N|$. Actually, these constraints are in correspondence with log-linear models over N.[5] Nevertheless, the list of conditions (C) is not reduced completely: some of these constraints are superfluous because they follow from the other ones combined with (A) and (B).[6] Moreover, each of the (C)-constraints can, owing to (A.2), be re-formulated equivalently in a kind of "standard" form

$$\sum_{S\in\mathcal{B}} k_S \cdot u(S) \leq 1 \quad \text{for } \mathcal{B} \subseteq \{S \subseteq N;\, |S| \geq 2\} \text{ and } k_S \in \mathbb{Z} \text{ for } S \in \mathcal{B}.$$

It looks like none of the constraints for $\mathcal{A} \subseteq \{S \subseteq N;\, |S| \geq 2\}$ is superfluous, while if $\mathcal{A}$ contains a singleton then both cases can occur: the respective inequality constraints can be either superfluous or non-derivable from others.[7]

Lemma 1. (the necessity of specific constraints)
If $u \in \mathsf{S}$ is a standard imset over N then the condition (C) is valid.

The proof is omitted because of limited scope of a conference contribution.

3.2 Conjecture about the linear constraints

The constraints (A)-(C) from the preceding section have several consequences, which are, perhaps, not evident at first sight. One of them is that every standard imset $u \in \mathsf{S}$ is bounded from below: $u(S) \geq -1$ for any $S \subseteq N$.

[5] This is because every class of sets closed under supersets is determined by the collection of its minimal sets, which is a class of incomparable sets. Hierarchical log-linear models also correspond to classes of incomparable subsets of the class $\{A \subseteq N; |S| \geq 1\}$, namely to those whose union is N.

[6] For example, if $\mathcal{A} = \{S \subseteq N;\ i \in S\}$ for some $i \in N$ then (A.2) gives $\sum_{S\in\mathcal{A}} u(S) = 0 \leq 1$.

[7] This happens in the case $|N| = 5$.

We have shown that (A)-(C) are necessary constraints on points in P, but we have also some reasons to conjecture that they are sufficient to characterize the standard imset polytope P. More specifically, we have verified for $|N| \leq 4$ that the conditions (A)-(C) characterize P. Thus, we dare to formulate the following hypothesis.

> *Conjecture* The linear constraints (A)-(C) together form a necessary and sufficient condition for $u \in \mathbb{R}^{\mathcal{P}(N)}$ to belong to P.

4 Lattice points in the standard imset polytope

Another related question concerning the polytope P is how "thick" it is. More specifically, we may ask whether there exists a lattice point in its interior. Raymond Hemmecke made some computations to find out whether such a point exists in the case $|N| \leq 5$ and the result was negative. This led him to a hypothesis that every lattice point in the standard imset polytope is already the standard imset. In this paper, we confirm the hypothesis:

Theorem 1. *If* $u \in \mathsf{P} \cap \mathbb{Z}^{\mathcal{P}(N)}$ *then* $u \in \mathsf{S}$.

The proof is quite technical and strongly depends on former results of ours [10]; specifically, it depends on the details of an algorithm for testing whether an imset is standard. It is omitted in this contribution.

In light of Theorem 1 one can formulate a weaker version of the conjecture from § 3.2:

> *Conjecture** The constraints (A)-(C) together form a necessary and sufficient condition for $u \in \mathbb{Z}^{\mathcal{P}(N)}$ to be a standard imset (over N).

Indeed, if *Conjecture* is true then, by Theorem 1, *Conjecture** holds as well. However, it is not clear at this moment whether the proof of *Conjecture** is enough to confirm the hypothesis from § 3.2.

5 Differential imsets over 4 variables

The result of our analysis of the geometric neighborhood in the case $|N| = 4$ is an electronic catalogue. To describe the catalogue we need a few auxiliary observations.

5.1 Some auxiliary concepts and results

Given a differential imset $w = u - v$ for $u, v \in \mathsf{S}$ it follows from the formula (1) that the *degree difference* $deg(u) - deg(v)$ does not depend on the choice of the pair $u, v \in \mathsf{S}$. This seems to be quite important characteristic of w.

We say that two imsets u, v over N are *permutation equivalent* if there exists a bijection $\pi : N \to N$ such that, for all $S \subseteq N$, it holds that $u(S) = v(\pi(S))$, where $\pi(S) = \{\pi(i);\, i \in S\}$. Each class of permutation equivalent imsets will be called a *PE class*. From the point of view of our analysis, it is not necessary to distinguish between permutation equivalent differential imsets. Every PE class can be described by an arbitrary representative.

Evidently, if $w = u - v$ is a differential imset for $u, v \in \mathsf{S}$ then $-w = v - u$ is a differential imset, too. Again, from the point of view of our analysis it is not necessary to distinguish between w and $-w$. Therefore, we keep only one of these in the catalogue. If the degree difference is non-zero we choose $w = u - v$ with $deg(u) > deg(v)$. That means, our catalogue only contains (PE representatives of) differential imsets with non-negative degree difference.

An important question is how to express the differential imsets. An elegant solution is offered below.

Lemma 2. *Every standard imset is a combination of elementary imsets with coefficients* $+1$ *(and* 0*).*

A kind of consequence of Lemma 2 is the following observation.

Lemma 3. *Every differential imset* $w = u - v$ *for* $u, v \in \mathsf{S}$ *is a combination of elementary imsets with coefficients* $+1$ *and* -1 *(and* 0*). Moreover, there exists a combination with at most* $\binom{|N|}{2}$ *non-zero coefficients.*

Proofs are omitted because of limited scope of the contribution. In particular, every differential imset for a pair of geometric neighbors can be expressed in that way, which we utilize in our catalogue.

5.2 Description of the catalogue

Our catalogue contains differential imsets $w = u - v$ for those $u, v \in \mathsf{S}$ that are *geometric neighbors*. It contains just one representative for each PE class and only imsets with non-negative degree difference are kept there.

We classified those differential imsets w using three criteria:

- the degree difference for w,
- the *squared Euclidean length* of w, that is, $\sum_{S \subseteq N} w(S)^2$, and
- the number of non-zero imset values, that is, $|\{S \subseteq N;\, w(S) \neq 0\}|$.

In the case $|N| = 4$ the degrees of differential imsets (for geometric neighbors) are integers from the interval $[0, 3]$. The values of the squared Euclidean length are even numbers from interval $[4, 22]$. The numbers of non-zero imset values are integers from interval $[4, 12]$.

There are 8518 ordered pairs (u, v) of geometric neighbors. As explained above, for each couple of ordered pairs (u, v) and (v, u), we have chosen only one differential imset out of $w = u - v$ and $-w = v - u$. In this way, we got 2831 differential imsets; they constitute 319 PE classes. Table 2 gives these numbers for each degree difference.

In order to understand better the geometric neighborhood we searched for an elegant description of differential imsets. One possible solution is offered by Lemma 3: every differential imset over 4 variables can be written as a combination (with coeficients $+1$ or -1) of at most 6 elementary imsets (out of 24 possible elementary imsets).

A complete catalogue of differential imsets over 4 variables with a detailed analysis for each differential imset is available at:

`http://staff.utia.cas.cz/vomlel/imset/catalogue-diff-imsets-4v.html`

Table 2: Numbers of geometric neighbor pairs, differential imsets and PE classes.

degree difference	neigh. pairs	diff. imsets	PE classes
0	2894	927	88
1	4248	1359	144
2	1296	505	71
3	80	40	16
total	8518	2831	319

5.3 A simple example

As mentioned in § 2.5, the classic inclusion neighborhood is contained in the geometric one and the inclusion neighbors are geometric neighbors with the degree difference ± 1.

One of our previous open questions was whether the converse holds. However, as one can deduce from Table 2, this is not true for $|N| = 4$: there are 144 PE classes with the degree difference 1 while one has only 3 PE classes of elementary imsets.

A simple example of a differential imset $w = u - v$ for geometric neighbors $u, v \in \mathsf{S}$ with degree difference 1 that is not an elementary imset is as follows:

$$w = \delta_{\{a\}} - \delta_{\{a,b\}} - \delta_{\{c,d\}} + \delta_{\{b,c,d\}},$$

where

$$u = \delta_{\emptyset} - \delta_{\{a,b\}} - \delta_{\{c,d\}} + \delta_{\{a,b,c,d\}}, \quad v = \delta_{\emptyset} - \delta_{\{a\}} - \delta_{\{b,c,d\}} + \delta_{\{a,b,c,d\}}.$$

6 Conclusions

Let us mention some of our research goals motivated by the results reported here. First, we would like either confirm or disprove the conjecture from § 3.2 for $|N| = 5$. If it is confirmed for $|N| = 5$ we may try to verify the weaker version of the conjecture from § 4 then.

The catalogue from § 5 is meant as a step towards a deeper analysis of the geometric neighborhood. For example, we would like to find out whether there is a graphical interpretation of geometric neighbors, namely whether differential imsets (for geometric neighbors) correspond to graphical operations with the respective essential graphs.

Acknowledgements

This research has been supported by the grants GAČR n. 201/08/0539 and MŠMT n. 1M0572. J. Vomlel has also been supported by the grants GAČR n. 2C06019 and Eurocores LogICCC n. ICC/08/E010.

References

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

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

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

[4] Neapolitan R.E. (2004) *Learning Bayesian Networks*, Pearson Prentice Hall.

[5] Schrijver A. (1986) *Theory of Linear and Integer Programming*, John Wiley.

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

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

[8] Studený M. (2005) *Probabilistic Conditional Independence Structures*, Springer-Verlag.

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

[10] Studený M., Vomlel J. (2009) A reconstruction algorithm for the essential graph, *International Journal of Approximate Reasoning* 50, 385-413.

[11] Studený M., Vomlel J., Hemmecke R. (2009) A geometric view on learning Bayesian network structures, accepted in *International Journal of Approximate Reasoning*.

Variable Selection in Local Regression Models via an Iterative LASSO

Diego Vidaurre
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
diego.vidaurre@fi.upm.es

Concha Bielza
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
mcbielza@fi.upm.es

Pedro Larrañaga
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
pedro.larranaga@fi.upm.es

Abstract

Locally weighted regression is a technique that predicts the response for new cases from their neighbors in the training dataset. In this paper we propose to join modern regularization approaches to locally weighted regression. Specifically, the LASSO method is able to select relevant variables leading to sparse models. We present two algorithms that embed LASSO in an iterative procedure that incrementally discard or add variables, respectively, in such a way that a LASSO-wise regularization path is locally obtained. The algorithms are tested in two different datasets from the UCI repository, obtaining promising results.

1 Introduction

Let $\boldsymbol{\chi} = \{\chi_1, ..., \chi_p\}$ denote the set of covariates and Y the response variable. Linear regression is a widely used tool concerning the influence of $\boldsymbol{\chi}$ over Y. This relationship is modelled by a linear combination of some of the covariates, such that a least squares function is minimized. Let $\mathscr{D} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), ..., (\boldsymbol{x}^{(n)}, y^{(n)})\}$ be the dataset containing the set of n points in the space of covariates and the response, where $\boldsymbol{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_p^{(i)})$. Let $\boldsymbol{X}$ denote the $n \times p$ matrix whose rows are the p-vectors $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(n)}$ and let $\boldsymbol{y} = (y^{(1)}, y^{(2)}, ..., y^{(n)})$ the vector of responses. Assuming the data is centered, the common linear regression model assumes a relationship such that:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ are the regression coefficients. Being $\boldsymbol{\Sigma}$ the $p \times p$ covariance matrix, the stochastic unobserved component $\boldsymbol{\epsilon}$ term is distributed:

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}). \tag{2}$$

Hence, there are p parameters to be determined, so that the sum of the squares of the distances from the response points to the line drawn by the linear equation is to be minimized. Typical hypothesis to be checked are linearity, normality, independence and variance homogeneity.

Since it bases its method on empirical loss minimization, linear regression may overfit the data. Regularization techniques add a penalization term to the usual regression preventing overfitting, reducing the variance of the estimates and giving rise to more interpretable models. Two widely used methods are *ridge* [1] and the *least absolute shrinkage regression and selection operator* [2]. We are focusing here on the second, commonly referred to as LASSO or l1-regularization. For a general review of LASSO, see [3]. A significant property of the LASSO is its ability to move many regression coefficients to zero, performing a variable selection (sparser models) at the same time than prediction. The LARS [4] algorithm is a variable selection and regression method that outperforms the classical forward stepwise regression algorithm [5], and solves the LASSO with a small modification in a very efficient way.

However, the response variable cannot be always predicted by means of a simple linear function of the covariates, and the results are not optimal from a statistical point of view. In this case some kind of nonlinear analysis may be required. In general, nonlinear regression procedures [6] intend to fit data to any selected equation, finding the values of the parameters that minimize the sum of the squares of the distances from the data points to the curve.

Sometimes, to perform a nonlinear analysis is not straightforward, and it is not possible to establish a unique function for the entire data space. In this case it is more convenient to use some form of local learning. A common method is the *locally weighted regression* (*LOESS*), built on classical least squares regression [7, 8, 9]. For each point in the covariate space, there is a neighborhood containing the point in which the regression surface is well approximated by a function from a parametric class. In this approach, instead of minimizing the residual sum of squares, a weighted sum of squares is minimized. The weights are provided by a function of the distances between the data and the point of interest, giving more importance to closer points. In [7] a second algorithm, called *robust locally weighted regression*, is proposed for providing robustness. In short, after firstly performing the LOESS procedure, the algorithm iteratively calculates new sets of weights basing on the residuals of the estimates $\hat{\mathbf{y}}$ regarding the real response $\mathbf{y}$, in such a way that large residuals correspond to small weights and vice versa. Regression and weights calculation are repeated until some stopping criterion is met. Also in the local fitting arena, in [10] the authors face nonlinearity by using a sum of smooth functions instead of a single parametric model for local learning.

A different form of local analysis is the spatial analysis. The *expansion method* [11] and the *geographically weighted regression* (GWR) [12] are well-known algorithms. Both assume that the influence of the covariates on the

response might vary according to the spatial location of the data, typically 2D coordinates where the data are collected. In the expansion method, the regression coefficients at a specific location are the result of a function of the location itself and a set of parameters (constant for all cases) to be learnt from the dataset. In the GWR algorithm, weights are locally assigned to data, so that nearer data are given more importance than further data.

There have been few attempts to combine local learning and variable selection with regularization. Also in the field of spatial analysis, the *geographically weighted LASSO* (GWL) [13] introduces a LASSO-wise penalization on the GWR estimated coefficients. Regardless of spatial analysis, ridge regression (along with principal components regression and partial least squares regression) is applied to local linear prediction of chaotic time series [14]. However, to the best of our knowledge, a LASSO penalization scheme for locally weighted regression has never been proposed.

Our contribution is a method based on LASSO both for local prediction and local variable selection. The setting is a scenario where usual linear regression is not appropriate, and a local approach seems to be convenient. We have developed two algorithms based on LARS for this aim. Unlike GWL, the distances are calculated in the covariate space instead of from separate location coordinates. A naïve approximation could be to add a LASSO penalty to the locally weighted regression. However we are using LASSO because we expect a sparse solution, and the irrelevant covariates should not have been used in the weights estimation (distance calculation). The problem lies in that the distance calculation is previous to the regression, and hence previous to know what variables are irrelevant. To overcome this obstacle, we suggest a couple of iterative algorithms that alternate variable selection with distance computation. One proceeds forwardly from the empty solution where all variables are excluded from the model and starts adding variables, one by one, until all the variables are in the model. The other works in the opposite backwards way, steming from the model with all variables and removing one by one until the empty solution is reached. At each step, distances are recalculated from the current variables in the model, assigning weights to the data for the following regularized regression. LARS is used for selection and removal of variables. As we will explain below, both local algorithms produce a pathway of solutions, from where a unique solution might be selected by means of some selection criterion.

The organization of the paper is as follows: Section 2 describes local regression and the LARS/LASSO algorithms in detail. Section 3 states the novel algorithms, that we are calling *forward local l1 selector*, and *backward local LARS selector*. Section 4 outlines the set of experiments to test the algorithms. Finally, in Section 5 we round the paper off with conclusions and future work.

2 Foundations

2.1 Local regression

The local regression method was originally devised for time series, where one expects that events close in time share common patterns. We focus on the LOESS procedure discussed in [7]. Although locally weighted regression paradigm is not limited to local linear fitting, we will not work in this paper with functions other

than linear. In [15] a bunch of mathematical properties of LOESS is discussed. [16] lists some advantages of using local regression.

Assuming the response to be centered, LOESS sets out the following local regression for $\boldsymbol{x}^{(k)}$:

$$\sum_{i=1}^{n}\left\{\left(y^{(i)}-\sum_{j=1}^{p}x_j^{(i)}\beta_j\right)^2 g\left(\frac{d(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(k)})}{\tau}\right)\right\}, \tag{3}$$

where $g(.)$ is a weight function, $d(.)$ is a distance function and τ is the bandwidth constant.

Let $\boldsymbol{w}^{(k)}=(w_1^{(k)},...,w_n^{(k)})$ be the vector of weights, with components

$$w_i^{(k)}=\sqrt{g\left(\frac{d(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(k)})}{\tau}\right)}, \tag{4}$$

and let $\boldsymbol{W}$ be the diagonal matrix whose elements $W_{ii}^{(k)}=w_i^{(k)}$, the vector of coefficients can be estimated as:

$$\boldsymbol{\beta}^{(k)}=[\mathbf{X}^T\boldsymbol{W}^{(k)T}\boldsymbol{W}^{(k)}\mathbf{X}]^{-1}\mathbf{X}^T\boldsymbol{W}^{(k)T}\boldsymbol{W}^{(k)}\boldsymbol{y}. \tag{5}$$

When a new point $\boldsymbol{x}^{(k)}$ comes up, its response $y^{(k)}$ is predicted by using ad-hoc coefficients $\boldsymbol{\beta}^{(k)}$ locally to the point itself and calculated just at this moment. The distribution of $y^{(k)}$ is unknown. This method is known as *lazy regression* [17]. If the procedure turns out to be too demanding for the abundant affluence of new $\boldsymbol{x}^{(k)}$, or we need a ready-to-use closed model for any reason, a possibility is to run the algorithm for each pair $(\boldsymbol{x}^{(i)},y^{(i)})$ and use for each new $\boldsymbol{x}^{(k)}$ the set of regression coefficients corresponding to the closer point in $\boldsymbol{X}$, say $\boldsymbol{x}^{(i1)}$. Depending on the distance from $\boldsymbol{x}^{(k)}$ to $\boldsymbol{x}^{(i1)}$, we can also decide either to calculate a new set of regression coefficients or to use $\boldsymbol{\beta}_{i1}$.

There are four relevant aspects when considering LOESS: the parametric family to be locally fitted, the fitting criterion, the weight function and the bandwidth [18].

As said above, we are focusing on the linear parametric family. Assuming $\mathbf{y}$ to be Gaussian with constant variance, least squares is a natural choice for the fitting criterion. If we cannot assure constant variance, some form of regularization may be used along with least squares. On the whole, the parametric family and the fitting criterion depend on the assumptions about the nature of the data and the distribution of the response. As we will detail in Section 3, we are using a penalized least squares favouring parsimony.

Regarding the weight function, any weight function that satisfies the properties listed in [7] may be used. The different choices we are using for the weight function are formulated in Section 3.

Finally, the choice of the bandwidth is a crucial parameter; the nature of the data, its cardinality and dimension, are relevant for a correct selection. On one hand, the bandwidth may be fixed beforehand or selected locally for each $\boldsymbol{x}^{(k)}$. The latter is particulary appropriate for online training [19] and yields some advantages in any case. Typically, it is done by a leave-one-out cross-validation, which can be solved recursively for an increased efficiency [20]. We have tested both fixed and variable bandwidth selection. However, in principle the recursive method of [20] cannot be applied here, as it is thought to solve the least squares

problem by the classical estimation method. On the other hand, the *curse of dimensionality* states that as far as the dimension p is bigger, the points quickly become sparse. In this case it is a good idea to increment the bandwidth to compensate this effect.

A relevant issue related to the extent of p is the adecuacy of local regression for high dimensionality. First, the analyst must take into account that local methods are relatively computationally intensive. The expected computation time for a LOESS estimate is the same than for a least squares fit, $O(p^3 + np^2)$, plus the complexity for the weights calculation, $O(np)$. For a single $\boldsymbol{x}^{(k)}$, n distances have to be calculated. It could be demanding, specifically when p is high. Furthermore, in principle, the estimated $\boldsymbol{\beta}^{(k)}$ is only valid for this point. If n is very large, there has been little work done for local regression. In [21] the author presents some validation tests to test the adecuacy of smoothing in binary logistic regression. Although the method and the scenario are slightly different, the conclusions are valid for the LOESS. In short, his analysis shows that the results are still reliable for increments of p if n is large enough, although the inclusion of irrelevant variables has a quite negative effect in the smoothing process. This is just the point we are tackling in this paper.

2.2 LARS/LASSO

During the last years, the original reference for the LASSO algorithm [2] has received over 1600 cites according to Google Scholar by the time this paper is being written. The LASSO estimates are defined as

$$\boldsymbol{\beta}_\alpha = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_j^{(i)} \beta_j)^2, \tag{6}$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq \alpha. \tag{7}$$

Unlike ordinary least squares and ridge regression, LASSO forces regression coefficients to become zero as we decrease the tunning parameter α. In this way it simultaneously performs variable selection and estimation. The complete solution of the LASSO for all values of α forms the *regularization path*. The regularization path usually starts with a small α and all coefficients equal to zero. One coefficient at a time is made different from zero, although from time to time any variable may also exit the model. For variable selection purposes, we only need to concern about a finite set of α values, specifically those that make the number of zero coefficients to change. Regarding estimation, we still need to pay attention only to this finite set, since the increments on the coefficients between two consecutive values of α are linear. This property makes the regularization path to be entitled as *piecewise constant*.

The LASSO is a quadratic programming problem with a linear inequality constraint. However, the LARS algorithm [4], designed for least angle regression, is able to calculate all possible LASSO estimates for a given problem in $O(p^3 + np^2)$ with a small modification. This is the same cost than for a usual least squares fit. In short, LARS is an iterative algorithm that starts with an empty

set of active (non-zero) variables and adds at each step one variable χ_t to this set. This is the one whose correlation with the residuals is the largest. The vector of correlations is:

$$\boldsymbol{c} = \boldsymbol{X}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}). \tag{8}$$

The coefficients of the variables in the active set are increased toward the direction of the least-squares fit based on such variables. So forth, a new variable gets active when its correlation with the residuals equals that of the active set.

Regarding the mathematical properties, there is an amount of theoretical work supporting the LASSO. For instance, in [22] consistency of LASSO is discussed and demonstrated under certain conditions. There have also been some variations of the original LASSO to improve such properties [23].

3 Local LARS/LASSO

In the discussion section, [8] comment the need of incorporating into the LOESS methodology a variable selection procedure when required, i.e, when we know of the presence of irrelevant variables. In this line of argument, we present two algorithms that combine l1-regularization with the usual locally weighted regression paradigm.

As commented above, a first possible approach is equivalent to the GWL algorithm in [13], that is, to directly apply a set of weights to the data set. The weights would be obtained from some transformation of the Euclidean or Mahalanobis distances to the point of interest. Whereas for GWL these distances come from separate coordinates, in locally weighted regression the distances are calculated in the space of covariates. Although simple and easy to implement, irrelevant variables are getting involved in the distance calculation task. Therefore, we state that this method is naive and ineffective, and it is expected to lead to incorrect predictions and incorrect feature selections. This effect will get more marked as the number of irrelevant variables increases. To simplify the terminology, we will call this method as *naive local selector*.

To minimize this risk we propose an iterative algorithm that calculates distances just on the active set of variables at each step. Two versions are presented: a forward algorithm and a backward algorithm.

The forward algorithm, that we will call *forward local l1 selector*, starts with an empty set of variables. It initializes a set of n weights on the distances over all variables. After appropriately weighting the data with them, it runs a LARS algorithm, stopping at the first iteration and keeping the first variable coming up. This variable, say indexed by j, will be the first member of the active set V. Then, weights are recalculated, but using only $V = \{\chi_j\}$, and LARS is run again over weighted data, that will stop when a variable not included in V appears. This variable is then included in V. It recalculates weights again, basing on the active set, and iterates like that until the active set contains all variables or any other stopping criterion is met. Note that at each step of the algorithm, LARS starts from zero variables and makes an arbitrary number of iterations, adding variables before reaching a variable not included in V. However, it is expected that, previous to get a variable not in V, LARS will pass through most of the variables in V at that moment.

Finally, when the algorithm completes all p iterations, some selection criterion is needed. Since the LARS method has run several times starting from zero variables, there are some solutions available with one variable, some solutions with two variables, etc. We will just select the best solution of each variable cardinality by the criterion applied to a separate test database. For example, as the algorithm iterates p times, and run LARS p times, there will be exactly p solutions with one variable. In this work we are using a minimum absolute error l1-penalized on the (weighted) test database. The pseudocode in Algorithm 1 roughly schematizes the procedure.

Algorithm 1 forward local l1 selector

Input: training data set $\boldsymbol{X}, \boldsymbol{y}$ with p variables and n cases,
Input: testing data set $\boldsymbol{X}', \boldsymbol{y}'$ with p variables and n' cases,
Input: bandwidth τ of the neighborhood,
Input: weight function $g(.)$ and distance function $d(.)$,
Input: point $\boldsymbol{x}^{(k)}$, whose response is to be predicted,
Output: set of coefficients $\boldsymbol{\beta}^{(k)}$
Calculate distances $\boldsymbol{d} = (d_1, d_2, ..., d_n)$ to $\boldsymbol{x}^{(k)}$ over all variables
$\boldsymbol{w} := g(\boldsymbol{d}, \tau)$ (vector of weights)
$\boldsymbol{W} := diagonal(\boldsymbol{w})$
$V := \{\}$ (active set)
$t := 0$
repeat
  $\boldsymbol{Xw} := \boldsymbol{W} * \boldsymbol{X}$ (weighted covariates)
  $\boldsymbol{yw} := \boldsymbol{W} * \boldsymbol{y}$ (weighted response)
  $\boldsymbol{paths}(t) := LARS(\boldsymbol{Xw}, \boldsymbol{yw})$ (stopping when a variable $\notin V$ appears)
  $V := V \cup \chi_j \quad | \quad \boldsymbol{path}_j(t) \neq 0$
  Calculate distances $\boldsymbol{d}$ to $\boldsymbol{x}^{(k)}$ using variables in V
  $\boldsymbol{w} := g(\boldsymbol{d}, \tau)$
  $\boldsymbol{W} := diagonal(\boldsymbol{w})$
  $t := t + 1$
until $|V| = p$
for $j := 1$ **to** p **do**
  $\boldsymbol{\beta}(j) = bestsolution(\boldsymbol{X}', \boldsymbol{y}', \boldsymbol{paths}, j)$,
  the best solution among those with j coefficients different from zero
end for

The second algorithm is a backward version of the forward local l1 selector, with some differences that we show straightaway. We call it the *backward local l1 selector*. The algorithm starts with all the p covariates, $V = \{\chi_1, \chi_2, ..., \chi_p\}$, calculates the weights and uses LASSO to discard one variable, say indexed by j. In a second step, it calculates the distances again on $V = V \backslash \chi_j$, and performs another LASSO regression with the new weights, discarding another variable. The algorithm keeps alternating variable selection with weights calculation until some stopping criterion is met, or until there are no variables left. Two possible stopping criteria are the similarity of the predictions and the similarity of the weights across iterations. In this work we are running the algorithm until all variables are run out. Note that a test database is not needed anymore, because LARS is run p times and only the last solution before stopping is kept. That

is, there is already only one solution for each variable cardinality. Hence, the entire dataset can be used for the training. We show the pseudocode for the backward local l1 selector in Algorithm 2.

Algorithm 2 backward local l1 selector

Input: training data set $\boldsymbol{X}, \boldsymbol{y}$ with p variables and n cases,
Input: bandwidth τ of the neighborhood,
Input: weighting function $g(.)$ and distance function $d(.)$,
Input: point $\boldsymbol{x}^{(k)}$, whose response is to be predicted,
Output: set of coefficients $\boldsymbol{\beta}^{(k)}$
$V = \{\chi_1, ..., \chi_p\}$ (active set)
Calculate distances $\boldsymbol{d}$ to $\boldsymbol{x}^{(k)}$ using variables in V
$\boldsymbol{w} := g(\boldsymbol{d}, \tau)$ (vector of weights)
$\boldsymbol{W} := diagonal(\boldsymbol{w})$
$\boldsymbol{Xw} := \boldsymbol{W} * \boldsymbol{X}$ (weighted covariates)
$\boldsymbol{yw} := \boldsymbol{W} * \boldsymbol{y}$ (weighted response)
$\boldsymbol{paths}(0) := LARS(\boldsymbol{Xw}, \boldsymbol{yw})$, taking the last solution, with all $\boldsymbol{\beta} \neq 0$
$t := 1$
repeat
  Calculate distances $\boldsymbol{d}$ to $\boldsymbol{x}^{(k)}$ using variables in V
  $\boldsymbol{w} := g(\boldsymbol{d}, \tau)$
  $\boldsymbol{W} := diagonal(\boldsymbol{w})$
  $\boldsymbol{Xw} := \boldsymbol{W} * \boldsymbol{Xv}$ (let $\boldsymbol{Xv} = \boldsymbol{X}$ but including only variables in V)
  $\boldsymbol{yw} := \boldsymbol{W} * \boldsymbol{y}$
  $\boldsymbol{paths}(t) := LARS(\boldsymbol{Xw}, \boldsymbol{yw})$, taking the last but one, with one $\beta_j = 0$
  $V = V \setminus \chi_j \quad | \quad \boldsymbol{paths}(t)_j = 0$
  $t := t + 1$
until $|V| = 0$

For both algorithms, we have used Euclidean distances, and for weighting we have employed the well-known tricube function used in [8]. Let $\boldsymbol{x}^{(k)}$ be the point that the local procedure concerns and $\boldsymbol{x}^{(i)}$ any other point. The tricube function establishes

$$w_i = \begin{cases} \left(1 - \left(\frac{d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(k)})}{d(\boldsymbol{x}^{(q)}, \boldsymbol{x}^{(k)})}\right)^3\right)^3 & \text{if } \boldsymbol{x}^{(i)} \in C_\tau \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $d(.)$ is a distance function, τ is the bandwidth, C_τ contains the τn closer points to $\boldsymbol{x}^{(k)}$ and $\boldsymbol{x}^{(q)}$ is the furthest point to $\boldsymbol{x}^{(k)}$ in C_τ.

Regarding the computation complexity of the algorithms, in principle we are running p times a LARS algorithm with complexity $O(p^3 + np^2)$, which would yield a complexity $O(p^4 + np^3)$ plus the distances calculation ($O(np)$). However, it is remarkable that for the backward version, as long as the algorithms iterates, LARS has to deal with fewer variables. Therefore the complexity is

$$\sum_{j=1}^{p} O(j^3 + nj^2) = \sum_{j=1}^{p} O(j^3) + \sum_{j=1}^{p} O(nj^2) = O\left(\left(\frac{j(j+1)}{2}\right)^2 + n\frac{j(j+1)(2j+1)}{6}\right). \tag{10}$$

For the forward version, LARS stops earlier, specially in the first iterations. The worst complexity is the same than for the backward algorithm, although the mean complexity is smaller (at iteration t, LARS will loop at most t times, but sometimes it stops before). Naive algorithm has the same complexity than LARS.

4 Experiments

In this section we face the algorithms with two real databases: *Housing* and *Forest Fires*. Both can be found in the UCI Repository [1]. Housing dataset deals with prices of housing in the suburbs of Boston. Besides the response variable (the price) it has 14 variables (integer and real), and 506 instances. Forest fires dataset, thoroughly described in [24], has 13 real attributes and 517 instances. It concerns the occurence of forest fires in the Montensinho natural park, Portugal. A logarithm function $ln(y_i + 1)$ has been applied on the response. Since among the variables we have the fires location coordinates, it is very suitable the use of some way of spatial analysis. However, we will abstain from including such analysis as it is not the concern of this work. Thus, we put the coordinates values into the independent variables set. The comparisons have been done with LASSO, usual LOESS and *Regression trees* (DT) [25].

Firstly, to compare local approaches, we have run a set of tests using constant bandwidths, experimenting with 12 values between 0.15 and 0.8. To choose the best solution of the pathway for the proposed algorithms, we have cross-validated with 1/4 of each dataset. For space reasons, we only show the results for the Housing dataset (see Figure 1). Tables 1 and 2 show also some useful statistics for Housing database. An equivalent table has not been shown for Forest fires database because the many small values of the responses (small prediction errors) are dominated by the few big values (big prediction errors). We have taken one by one all points in the datasets, we have predicted their responses and compared to the true response. For each bandwidth and each algorithm, we show mean prediction errors and the mean numbers of variables. The best solution of each pathway has been selected with a l1-penalized score, basing on a separate test proportion of the database. Figure 1 shows that for small bandwidths the performance of the proposed algorithms is better than LOESS. However, LASSO turns out to be more accurate excepting for big bandwidths of LOESS. It reveals that the dataset can be up to a point linearly approximated. The good news is that the local approach needs some less variables to make the prediction.

Besides a competitive performance, the proposed algorithms seem to behave more robustly. It can be observed in the worst case (maximum error) and standard deviation for LOESS, LASSO and Regression Tree, significantly bigger than for the other algorithms.

As mentioned above, in local regression it is common to use an ad-hoc bandwidth for each point to be predicted. To analyze this method, in Figure 2 split prediction errors and number of variables in 15 intervals and plot histograms for each algorithm, for the Forest fires dataset. Again, we have used cross-validation with 1/4 of the dataset. We have cut off errors above 3.0. This is done to make

[1] http://archive.ics.uci.edu/ml/

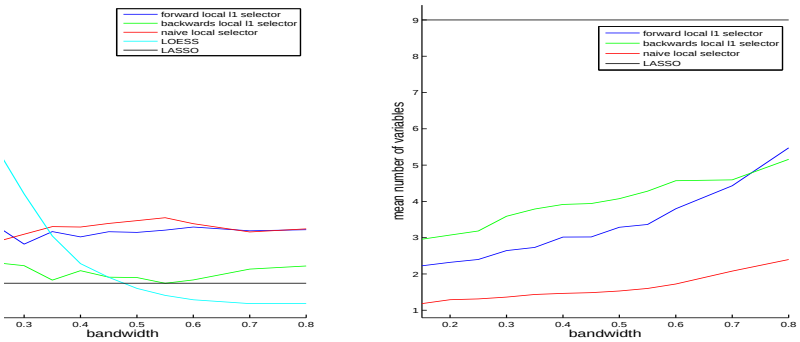


Figure 1: Evolution of the mean error and the mean number of variables for an increasing bandwidth, for Housing dataset. For LOESS, the number of variables is always the maximum, so it has been omitted.

	Ff	Bf	Nf	LRf	Fv	Bv	Nv	LRv	L	RT
mean	3.8	3.6	3.9	4.3	4.3	4.2	4.0	3.4	3.5	2.9
median	2.5	2.4	2.7	2.0	2.8	2.7	2.5	1.9	2.3	1.9
std dv.	4.4	3.9	4.1	7.3	4.9	4.7	4.4	5.7	3.8	3.6
max.	27.9	23.8	26.8	93.3	27.4	27.2	27.5	61.5	30.8	30.7

Table 1: Some statistics for estimation error, for Housing dataset, and algorithms: forward local l1 selector and fixed bandwidth (Ff), backward local l1 selector and fixed bandwidth (Bf), naive local l1 selector and fixed bandwidth (Nf), LOESS and fixed bandwidth (LRf), forward local l1 selector and adaptive bandwidth (Fv), backward local l1 selector and adaptive bandwidth (Bv), naive local l1 selector and adaptive bandwidth (Nv), LOESS and adaptive bandwidth (LRv), classical LASSO (L) and Regression tree (RT). For fixed bandwidth cases, a value of 0.3 has been taken.

	Ff	Bf	Nf	Fv	Bv	Nv	L
mean	2.6	3.6	1.4	2.1	2.8	1.2	9.0
median	2.0	3.0	1.0	2.0	3.0	1.0	9.0
std dv.	1.8	1.9	1.0	1.4	1.5	0.6	0.0
max.	10.0	12.0	7.0	8.0	7.0	6.0	9.0

Table 2: Same setting than for Table 1, for average number of variables. Local Regression algorithms and Regression tree have been removed.

intervals more specific and informative, ignoring outliers. This is fair play anyway with regard to our comparisons, since most big errors were obtained with LASSO and Regression trees. As observed, results are slightly better for the proposed algorithms than for LOESS and LASSO. Specifically, for LASSO most errors concentrate around 1, whereas for the other algorithms many error are smaller. It is remarkable the good behaviour of the Regression tree in this case, although it produces more big errors than the proposed algorithms. Note that the best solution for the naive version and LASSO approach often yields all regression coefficients equal zero, which obviously generalizes poorly. This is so because there are many zero responses in this dataset. However, the forward and backward versions usually recover at least two variables.

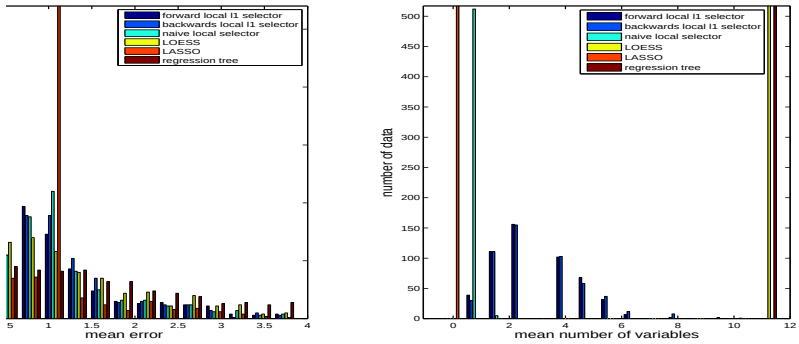


Figure 2: Histogram of 15 intervals of error and number of variables, for Forest fires dataset.

[24] emphasizes the importance of the prediction of small fires, which are the great majority. In Figure 5 we show a scatter plot of the response against the error. We exclude backward and naive approaches because of its similarity with the forward approach. All local algorithms were run with an adaptable bandwidth. The proposed algorithms are the ones which best predicts small fires, although all the methods have certain difficulties with responses equal to zero.

5 Discussion

In this work, we propose two variable selection and shrinkage iterative methods that lean on traditional locally weighted regression paradigm and l1-regularization. We prove its usefulness in two real datasets from the UCI repository, but we feel that better results are possible. The nature of the data is important to decide the adecuacy of the proposed methods. Specifically, the methods would stand out when the relation between covariates and response is sparse and nonlinear.

From the regularization side, we are providing an alternative for dealing with nonlinear data. From the local regression side, we supply the variable selection functionality. Moreover, using regularization techniques we can overcome the

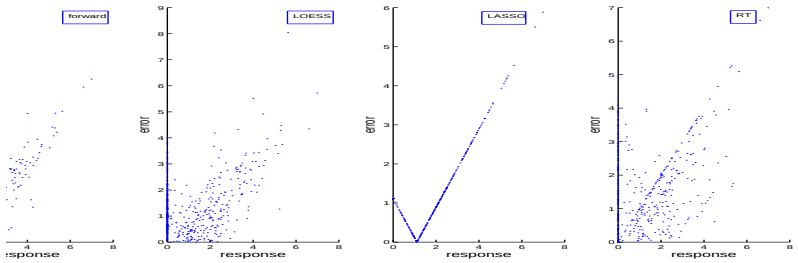


Figure 3: Scatter plots for Forest fires dataset, for forward local l1 selector, LOESS, LASSO and Regression tree, respectively. Responses are not centered

$p >> n$ case, that is, when the data matrix is not invertible As a derivation of least squares regression, locally weighted regression needs the cardinality of the dataset to be greater or equal than the number of variables.

Our approach is lazy in the sense that we lack an overall model valid for all future cases. Hence, as happens with locally weighted regression, we need to run the whole algorithm each time a new case is presented. Flexibility against nonlinearity and best performance of prediction are the advantages gained in exchange for a more expensive computation if compared with linear techniques. Although the way we are proceeding here is lazy, if computation time is a main concern, the analysis can draw the regression coefficients for some or all the cases in the training dataset, and extrapolate the new case to the closest points in the dataset in some way (for example giving a weighted mean of the "closest" responses).

Future work will revolve around the adaptation of the algorithms to multiresponse problems, applications to challeging data, use of recent variations of LASSO, and improvements over the algorithms. Specifically, we expect to develop a more sophisticated method for the selection phase of the forward algorithm. We consider this algorithm the most sensible and promising, but it needs to set aside a piece of the dataset to sieve solutions. This is a disadvantage against backward and naive algorithms that claims to be solved. Robustness is also an important concern. There are robust versions that prevent the bad effects of outliers both for LOESS [7] and for LASSO [26]. Methods that make the proposed algorithms more robust need to be investigated.

References

[1] A. Hoerl and R. Kennard, "Ridge regression: Biased estimates for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[2] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, pp. 267–288, 1996.

[3] T. Hesterberg, N. M. Choi, L. Meier, and C. Fraley, "Least angle and l_1 penalized regression: A review," *Statistics Surveys*, vol. 2, pp. 61–93, 2008.

[4] B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32(2), pp. 407–499, 2004.

[5] S. Weisberg, *Applied Linear Regression*. Wiley, New York, 1980.

[6] G. Seber and C. Wild, *Nonlinear Regression*. Wiley, New York, 1989.

[7] W. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74(368), pp. 829–836, 1979.

[8] W. Cleveland and S. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83(403), pp. 596–610, 1988.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Predictions*. Springer Verlag, New York, 2001.

[10] T. Hastie and R. Tibshirani, "Generalized additive models," *Statistical Science*, vol. 1(3), pp. 297–310, 1986.

[11] J. Jones and E. Casetti, *Applications of the Expansion Method*. Routledge, 1992.

[12] A. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, 2002.

[13] D. C. Wheeler, "Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted Lasso," *Environment and Planning A*, vol. 41, pp. 722–742, 2009.

[14] D. Kugiumtzis, O. C. Lingjaerde, and N. Christophersen, "Regularized local linear prediction of chaotic time series," *Physica D: Nonlinear Phenomena*, vol. 112, pp. 344–360, 1998.

[15] S. Devlin, "Locally-weighted multiple regression: Statistical properties and its use to test for linearity," Bell Communications Research, Piscataway, NJ, Tech. Rep., 1986.

[16] T. Hastie and C. Loader, "Local regression: Automatic kernel carpentry," *Statistical Science*, vol. 8(2), pp. 120–143, 1993.

[17] D. Aha, "Special issue on lazy learning," *Artificial Intelligence Review*, vol. 11(1-5), pp. 1–6, 1997.

[18] W. Cleveland and C. Loader, "Smoothing by local regression: Principles and methods," *Statistical Theory and Computational Aspects of Smoothing*, pp. 10–49, 1996.

[19] G. Bontempi, M. Birattari, and H. Bersini, "Lazy learning for local modelling and control design," *International Journal of Control*, vol. 72(7), pp. 643–658, 1999.

[20] M. Birattari, G. Bontempi, and H. Bersini, "Lazy learning meets the recursive least squares algorithm," in *Advances in Neural Information Processing Systems 11*. MIT Press, 1999, pp. 375–381.

[21] E. Fowlkes, "Some diagnostics for binary logistic regression via smoothing," *Biometrika*, vol. 74(3), pp. 503–515, 1987.

[22] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Machine Learning Research*, vol. 7, pp. 2541–2567, 2006.

[23] H. Zou, "The adaptive Lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101(12), pp. 1418–1429, 2006.

[24] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data," in *Portuguese Conference on Artificial Intelligence*, 2007, pp. 512–526.

[25] L. Breiman, J. Friedman, R. Ohlsen, and C. Stone, *Classification and Regression Trees*. Wadsworth, Monterey, CA, 1984.

[26] J. Khan, S. V. Aelst, and R. Zamar, "Robust linear model selection based on least angle regression," *Journal of the American Statistical Association*, vol. 102(480), pp. 1289–1299, 2007.

An Experimental Comparison of Triangulation Heuristics on Transformed BN2O Networks*

Jiří Vomlel
Institute of Information Theory and Automation of the AS CR,
Academy of Sciences of the Czech Republic
http://www.utia.cas.cz/vomlel

Petr Savicky
Institute of Computer Science
Academy of Sciences of the Czech Republic
http://www.cs.cas.cz/savicky

Abstract

In this paper we present results of experimental comparisons of several triangulation heuristics on bipartite graphs. Our motivation for testing heuristics on the family of bipartite graphs is the rank-one decomposition of BN2O networks. A BN2O network is a Bayesian network having the structure of a bipartite graph with all edges directed from the top level toward the bottom level and where all conditional probability tables are noisy-or gates. After applying the rank-one decomposition, which adds an extra level of auxiliary nodes in between the top and bottom levels, and after removing simplicial nodes of the bottom level we get so called BROD graph. This is an undirected bipartite graph. It is desirable for efficiency of the inference to find a triangulation of the BROD graph having the sum of table sizes for all cliques of the triangulated graph as small as possible. From this point of view, the minfill heuristics perform in average better than other tested heuristics (minwidth, h1, and mcs).

1 Introduction

A BN2O network is a Bayesian network having the structure of a bipartite graph with all edges directed from the top level toward the bottom level and where all conditional probability tables are noisy-or gates. Let $U = \{u_1, \ldots, u_m\}$ be the nodes of the top level of a BN2O network and $V = \{v_1, \ldots, v_n\}$ be the nodes of the bottom level of this network.

*J. Vomlel was supported by grants number 1M0572 and 2C06019 (MŠMT ČR), ICC/08/E010 (Eurocores LogICCC), and 201/09/1891 (GA ČR). P. Savicky was supported by grants number 1M0545 (MŠMT ČR), 1ET100300517 (Information Society), and by Institutional Research Plan AV0Z10300504.

In order to perform efficient inference, we transform these networks using tensor rank-one decomposition [4, 11, 7]. The rank-one decomposition (ROD) graph [8] of a BN2O graph G is the undirected graph constructed from G by

- adding an auxiliary node w_i for each $v_i \in V$,
- replacing each directed edge (u_j, v_i) by an undirected edge $\{u_j, w_i\}$, and
- adding an undirected edge $\{v_i, w_i\}$ for each $v_i \in V$.

The ROD graph is further transformed by triangulation resulting in an undirected triangulated graph. Note that nodes $v_i \in V$ are simplicial in the ROD graph and have degree one. Therefore we can perform optimal triangulation of the ROD graph by optimal triangulation of its subgraph induced by nodes $U \cup W$ [1]. This graph will be called the BROD graph [8]. See Figure 1 for an example of the BROD graph.

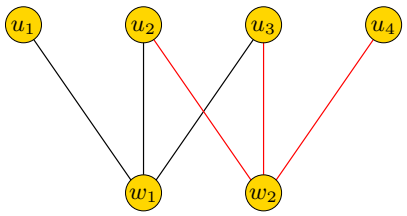


Figure 1: An example of the BROD graph

An important parameter for the inference efficiency is the total table size after triangulation. The table size of a clique C in an undirected graph is $\prod_{v \in C} |X_v|$, where $|X_v|$ is the number of states of a variable X_v corresponding to a node v. If all variables are binary the table size of a clique C is $2^{|C|}$. The total table size of a triangulation is defined as the sum of table sizes for all cliques of the triangulated graph. Therefore, it is desirable to find a triangulation of the BROD graph having the total table size as small as possible. Since this problem is known to be NP-hard and remains NP-hard for bipartite graphs [2], diferent heuristics are often used.

In this paper we perform experimental comparisons of existing heuristic triangulation methods applicable to the BROD graph, which is an undirected bipartite graph. This extends the results already published in [8]. Let us point out that the class of all possible BROD graphs is the same as the class of all bipartite graphs. We talk about BROD graphs, since this corresponds to our motivation.

2 Triangulation heuristics

In Section 3 we will experimentally compare triangulation heuristics minfill [6], minwidth [6], maximum cardinality search [10], and h1 [3]. In order to describe these heuristics, we need notions defined below.

Definition 2.1 Let $G = (V, E)$ be an undirected graph and $U \subseteq V$. The subgraph of G induced by a set of nodes U, denoted $G[U]$, is $G[U] = (U, F)$, where $F = \{\{u, v\} \in E : u, v \in U\}$.

Definition 2.2 Let $F(v) = \{\{v_1, v_2\} : \{v_1, v\} \in E, \{v_2, v\} \in E\}$.

In Table 1 we describe a general template for the considered triangulation heuristics except of minimum cardinality search. The criterion $\phi(u)$ used in step 1 in the template is different for different heuristics and is as follows.

Definition 2.3 *Let v be a node a a graph $G = (V, E)$. Then, let*

1. *$\phi_{minfill}(v)$ be the number of edges added if v is chosen, i.e., $\phi_{minfill}(v) = |F(v) \setminus E|$.*
2. *$\phi_{minwidth}(v)$ be the degree of v, $\phi_{minwidth}(v) = |nb_G(v)|$.*
3. *$\phi_{h1}(v)$ be the size of the largest clique containing $nb_H(v)$, where H is the induced subgraph of $(V, E \cup F(v))$ on the set $V \setminus \{v\}$.*

Table 1: General template for triangulation heuristics using criterion ϕ

For $i = 1, \ldots, |V|$ do:

1. Select a node v of graph G as $v = \arg\min_{u \in V} \phi(u)$, breaking ties arbitrarily.
2. Set $f(v) = i$.
3. Make v a simplicial node in G by adding edges to G, i.e., $G = (V, E \cup F(v))$.
4. Eliminate v from the graph G, i.e. replace G by $G[V \setminus \{v\}]$.

Return f.

Maximum cardinality search has slightly different structure than previously described heuristics. See Table 2.

Table 2: Maximum cardinality search

For all $v \in V$ set weight $w(v) = 0$.
For $i = |V|, \ldots, 1$ do:

1. Select an unnumbered node v of graph G maximizing weight w, breaking ties arbitrarily.
2. Set $f(v) = i$.
3. For all unnumbered nodes $u \in nb_G(v)$ set $w(u) = w(u)+1$.

Return f.

3 Experiments

We performed an experimental comparison of the triangulation heuristics on three types of random BN2O graphs. In the first set of experiments, we used 1300 BN2O networks, whose edges were chosen from the uniform distribution on all edges of a complete directed bipartite graph of a given dimension. In the second and third set of experiments we used submodels of the decision theoretic version of Quick Medical Reference (QMR-DT) model using a determinisitic choice of the nodes at the top level and two different types of random choice of the nodes at the bottom level. We will call these submodels QMR thumbnails.

3.1 Randomly generated BN2O networks

First, similarly to [8], we compared the triangulation heuristics on 1300 BN2O networks randomly generated with varying values of the following parameters:

- x, the number of nodes on the top level,
- y, the number of nodes on the bottom level, and
- e, the average number of edges per node on the bottom level.

For each x-y-e type, $x, y = 10, 20, 30, 40, 50$ and $e = 3, 5, 7, 10, 14, 20$ (excluding those with $e \geq x$) we generated randomly ten BN2O graphs by choosing the set of edges from the uniform distribution on the set of all e-tuples of edges from the $x \cdot y$ edges of the complete bipartite graph.

3.2 QMR-DT thumbnails

The decision theoretic version of the Quick Medical Reference [9] (abbreviated QMR-DT) is a large Bayesian network version of the original Quick Medical Reference [5]. There are 570 diseases and 4075 observations in the model. The structure of the model is a directed bipartite graph with edges directed from diseases in the top level to observations in the bottom level. All variables are binary and conditional probability tables of observations given diseases are noisy-or gates. Therefore, QMR-DT represents an example of BN2O model.

Testing triangulation heuristics on the whole QMR-DT is very time consuming and the analysis of this model requires specific algorithms. Our goal is to test the heuristics on smaller graphs. However, we want to test the heuristics on graphs, which contain substructures similar to those, which may appear in real applications. For this purpose, we split the top level of QMR-DT into 10 or 20 disjoint intervals of indices in the order of the nodes, in which the model is presented. This choice implies that similar nodes have higher chance to be chosen to the same subgraph. The exact bounds of the k intervals, where $k = 10$ or $k = 20$, were computed as $[s_{i-1} + 1, s_i]$, where $i = 1, \ldots, k$ and $s_i = \lfloor 570 \cdot i/k \rfloor$.

For each of the k intervals in the top level, denoted X, we used two types of random selection of the set Y of $y = \lceil 4075/k \rceil$ nodes in the bottom level and generated 10 randomly selected sets Y using each of the two methods. Hence, each interval X yields 20 pairs (X, Y) describing a submodel of QMR-DT of the required size. We used the following two types of random selection of Y.

- **Selection by edges.** We choose a random permutation of the edges with the starting point in X from the uniform distribution on such permutations and consider the sequence of the end points of these edges. Then, Y is the set of the first y different nodes in this sequence.
- **Selection by nodes.** We consider the set of end points of the edges, whose starting point is in X. Then, Y is a random subset of these end nodes of size y chosen from the uniform distribution on such subsets.

When $k = 10$, we obtain 200 models, which form the group of thumbnails denoted QMR-DT-57-408. When $k = 20$, we obtain 400 models, which form the group denoted as QMR-DT-29-204.

3.3 Results of experiments

Triangulation heuristics were tested on the BROD graphs G_{BROD}. We used the total table size tts of the graph G^h_{BROD} triangulated by a triangulation heuristics h as the criterion for comparisons. We used the *minfill* method as the base method against which we compared all other tested methods. Since randomness is used in the triangulation heuristics we run each heuristics ten times on each model and selected a triangulation with the minimum value of total table size tts.

For each tested model we computed the decadic logarithm ratio

$$r(h, minfill) \quad = \quad \log_{10} tts\left(G^h_{BROD}\right) - \log_{10} tts\left(G^{minfill}_{BROD}\right) \quad ,$$

where h stands for the tested triangulation heuristics.

We used three sets of models for the experiments:

- 1300 randomly generated models *x-y-e* from Section 3.1,
- 200 larger QMR thumbnails *QMR-DT-57-408* from Section 3.2, and
- 400 smaller QMR thumbnails *QMR-DT-29-204* from Section 3.2.

For each of these three groups of models we computed the tts estimate produced by heuristics $h \in \{minfill, minwidth, mcs\}$. For groups x-y-e and QMR-DT-29-204, we additionally computed the triangulation by $h = h1$. The obtained values of tts for $h \neq minfill$ were than compared to the results of $minfill$ for the same group of models. We eliminated the pairs of values of tts for h and $minfill$, which are equal, and performed two-sided Wilcoxon two-sample tests of the null hypothesis that the distribution of $r(h, minfill)$ is symmetric about 0 on the cases, where the two heuristics produced different values. The alternative hypothesis is that the distribution of $r(h, minfill)$ is biased towards negative or positive values. In order to asses, which sign of the typical difference is more likely, we present not only the p-values of the test, but also the values of the statistics W_+ and W_-. If $W_+ > W_-$, then the tested statistics is typically worse than minfill, when $W_+ < W_-$, then it is typically better. The results are summarized in Tables 3, 4, and 5, where nr. obs. means the number of models (observations), for which h and $minfill$ yield different tts.

The tests revealed that *minfill* performs significantly better than *mcs* on all three sets of models.

Table 3: Results of Wilcoxon test for models x-y-e

h	nr. obs.	W_+	W_-	p-val
minwidth	486	78150	40191	8.96e−10
mcs	1266	802011	0	0.00e+00
h1	499	87367	37383	8.88e−15

Table 4: Results of Wilcoxon test for models QMR-DT-57-408

h	nr. obs.	W_+	W_-	p-val
minwidth	193	12946	5775	3.95e−06
mcs	200	20100	0	0.00e+00

Also, *minfill* performed significantly better than *minwidth* on the set of randomly generated models *x-y-e* and on the model set *QMR-DT-57-408*, while on the model set *QMR-DT-29-204* the difference was not significant. On the model set *x-y-e* the advantage of *minfill* over *minwidth* increases with larger value of *tts*, which was not observed on the other test sets, see Figure 2.

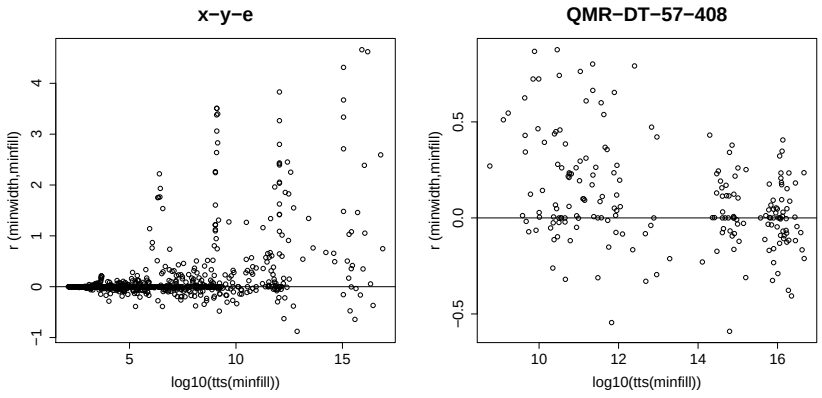


Figure 2: Dependence of $r(minwidth, minfill)$ on decadic logarithm of *tts* of *minfill* for the set of randomly generated models *x-y-e* and on the model set *QMR-DT-57-408*.

The computations of the *h1* heuristics on *QMR-DT-57-408* took too long, which kept us from the comparisons of *minfill* with *h1* on this model set. On the set of randomly generated models *x-y-e* *minfill* performed significantly better than *h1* heuristics, while on the model set *QMR-DT-29-204* the difference was not significant.

In Figures 3, 4 and 5 we present histograms of values of $r(h, minfill)$ for *x-y-e*, *QMR-DT-29-204*, and *QMR-57-408* model sets.

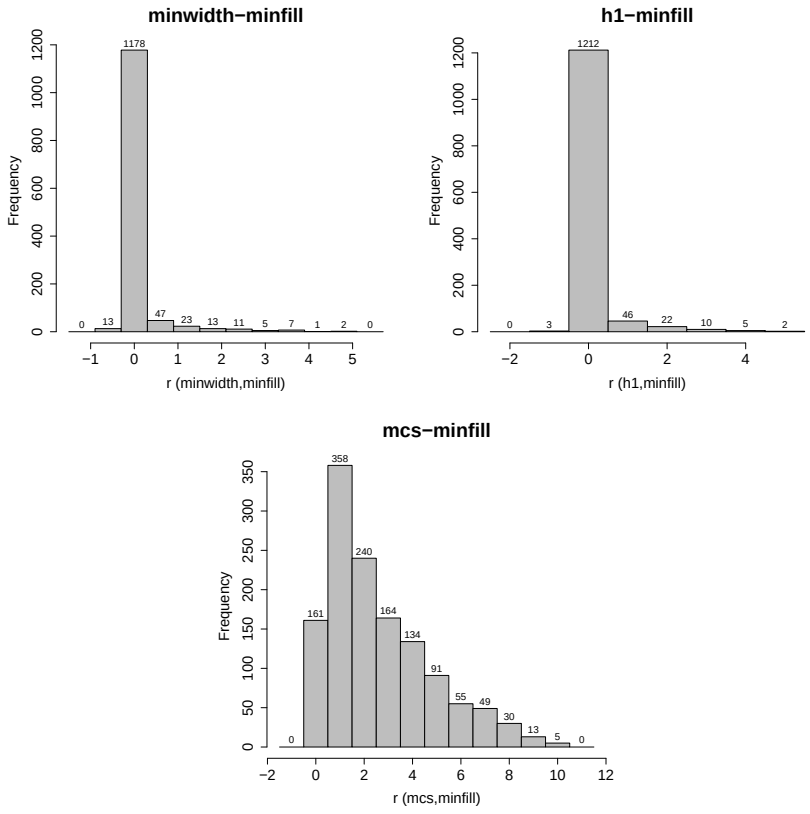


Figure 3: Histograms of values of $r(h, minfill)$ for x-y-e.

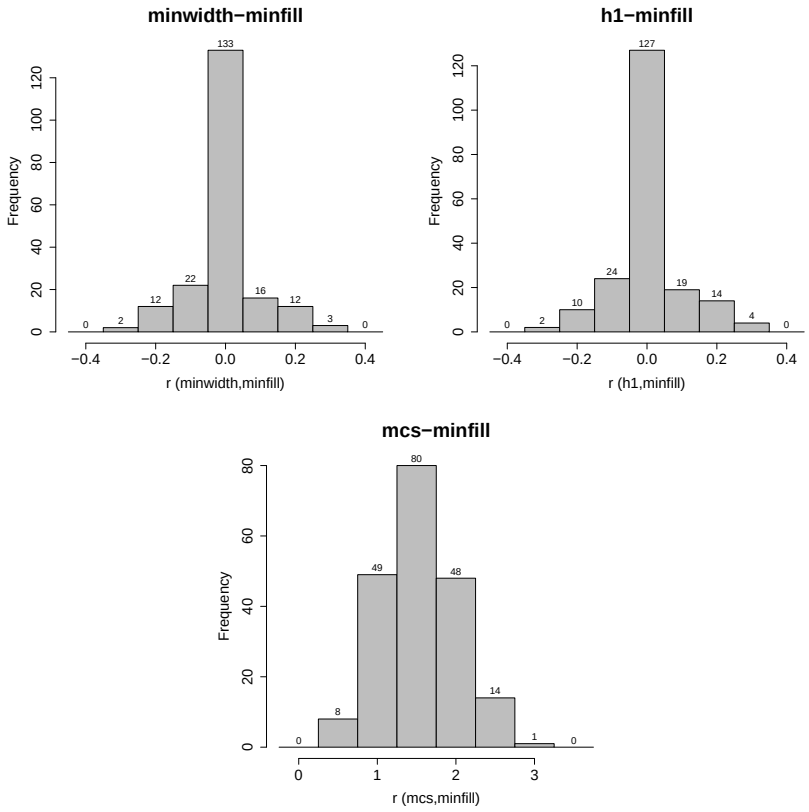


Figure 4: Histograms of values of $r(h, minfill)$ for *QMR-DT-29-204*.

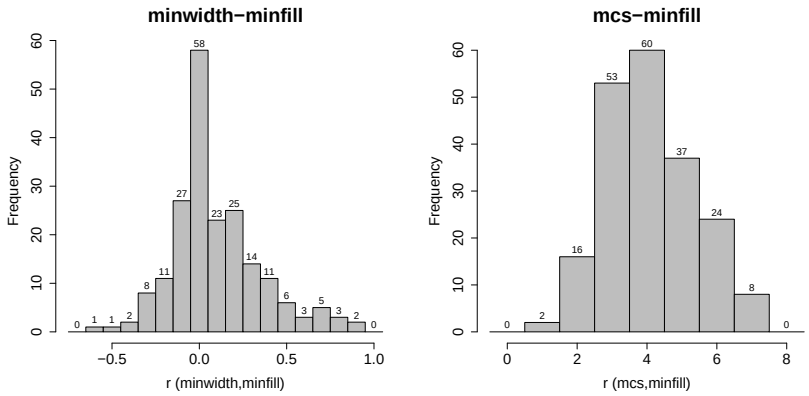


Figure 5: Histograms of values of $r(h, minfill)$ for *QMR-DT-57-408*.

Table 5: Results of Wilcoxon test for models QMR-DT-29-204

h	nr. obs.	W_+	W_-	p-val
minwidth	313	24517	24624	0.9736
mcs	400	80200	0	0.0000
h1	325	32008	20967	0.0011

4 Conclusions

In this paper we presented results of experimental comparisons of existing heuristic triangulation methods applicable to the BROD graph. The results of experiments reveal that, although no heuristics was dominant on all graphs, in average, the minfill heuristics gave the best results from the tested heuristics.

References

[1] H. L. Bodlaender, A. M. C. A. Koster, and F. Van Den Eijkhof. Preprocessing rules for triangulation of probabilistic networks. *Computational Intelligence*, 21(3):286–305, 2005.

[2] Hans L. Bodlaender and Fedor V. Fomin. Tree decompositions with small cost. *Discrete Applied Mathematics*, 145(2):143–154, 2005.

[3] A. Cano and S. Moral. Heuristic algorithms for the triangulation of graphs. In B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh, editors, *Advances in Intelligent Computing – IPMU '94: Selected Papers*, pages 98–107. Springer, 1994.

[4] F. J. Díez and S. F. Galán. An efficient factorization for the noisy MAX. *International Journal of Intelligent Systems*, 18:165–177, 2003.

[5] R. A. Miller, F. E. Fasarie, and J. D. Myers. Quick medical reference (QMR) for diagnostic assistance. *Medical Computing*, 3:34–48, 1986.

[6] D. J. Rose. A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations. *Graph Theory and Computing*, pages 183–217, 1972.

[7] P. Savicky and J. Vomlel. Exploiting tensor rank-one decomposition in probabilistic inference. *Kybernetika*, 43(5):747–764, 2007.

[8] P. Savicky and J. Vomlel. Triangulation heuristics for BN2O networks. In C. Sossai and G. Chemello, editors, *Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2009)*, 2009.

[9] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255, 1991.

[10] R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.*, 13:566–579, 1984.

[11] J. Vomlel. Exploiting functional dependence in Bayesian network inference. In *Proceedings of the 18th Conference on Uncertainty in AI (UAI)*, pages 528–535. Morgan Kaufmann Publishers, 2002.

Divergence Weighted Independence Graphs for the Multivariate Analysis of Survey Data

Joe Whittaker
Department of Mathematics and Statistics
Lancaster University
joe.whittaker@lancaster.ac.uk

Abstract

The analysis of survey data, collected on a set of response variables defined over a finite population, benefits from a bird's eye view of their inter-relationships and in particular, of their strengths. This overall analysis should highlight those variables that strongly modify the conditional distribution of another variable, and by contrast, should indicate those which have little affect. The weighted graph based on divergence measures of independence strength calculated from the sample fulfills this purpose. Survey data from the 1970 British Cohort Study provides an example for this methodology.

1 Introduction

Whittaker and Kao (2009) introduce the divergence weighted independence graph (dwig) to give a high level overview of dependency between categorical variables. The dwig gives a visual representation of the strengths of association with respect to a specified collection of observed survey variables. The weights are measures of mutual information between categorical variables in given marginal, joint and conditional distributions, and are measured in the common unit of millibits. In the next section a brief outline of the principal concepts of dwigs is given, which is followed by an extended application to the 1970 British Cohort Study.

2 Outline of the theory

We define a dwig to portray the independence relationships manifest in a set of variables $Y_1, Y_2, \ldots, Y_k$. These relationships are defined entirely in terms of population measures and are then estimated using their sample equivalents. A k-dimensional divergence weighted independence graph is the graph $G = (V, E, W)$, with vertices $V = \{i | i = 1, \ldots, k\}$, all edges $E = \{(i, j) | i, j \in V\}$ and weights $W = \{w_{ij} | i, j \in V\}$. We may define several dwigs to focus on different aspects of the joint distribution of variables in V.

In any one dwig each edge (i, j) corresponds to a single pairwise conditional independence statement between the random variables corresponding to its vertices. The weight w_{ij} is the information divergence corresponding to that statement, and is a measure of edge strength. A natural form of display is to set the width and tone of each edge proportional to the or edge strength. Each graph is complete.

The difference between different dwigs is how the array of conditioning sets is chosen, and this determines (by convention) which edges are undirected or directed. These arrays have the property that, in some sense, they specify the joint distribution of the variables in V. If all of the weights are zero so that all the corresponding independence statements hold, then the resulting graph should be the graph of mutual independent variables. If a subset of the weights are zero so that the corresponding independence statements hold, then the resulting graph is the graph of a proper distribution.

The weights of the *undirected divergence weighted conditional independence graph* are

$$w_{ij} = \inf(Y_i \perp\!\!\!\perp Y_j | Y_{V\setminus\{i,j\}}),$$

specified by conditioning on the subset $V\setminus\{i, j\}$, known as the 'rest', so w_{ij} is the extra information for predicting Y_i provided by Y_j after conditioning upon the rest. The information measure is the mutual information, see Whittaker (1990); Cover and Thomas (2002). If the edge (i, j) is excluded from E when $w_{ij} = 0$ and all other weights are set equal 1, the graph is identical to the classic conditional independence graph of Darroch et al. (1980).

There are two advantages of using conditional measures of mutual information of the form $\inf(Y_i \perp\!\!\!\perp Y_j | Y_{V\setminus\{i,j\}})$. Firstly the resulting graph approximates the conditional independence graph and makes its separation or Markov properties available, Lauritzen (1996). Secondly using the conditional divergence implies the strong neighbours of any vertex (those with high edge strengths) are always required for the best prediction of the vertex.

Chain graphs, Wermuth and Cox (1996), incorporate the use of both directed and undirected edges. It is assumed that there is a partial ordering, $<$, on the vertex set so that V can be partitioned into m subsets or *blocks* and the blocks form a chain $b_1 < b_2 < \ldots < b_m$. For instance, the variables in b_1 are potential parents of the variables in b_2, and the variables in $b_1 \cup b_2$ are potential parents of the variables in b_3, and so on. All directed edges connect variables from different blocks and are directed away from the preceding block. The *divergence weighted conditional independence chain graph* has two types of edges: directed and undirected, with weights defined as follows. All undirected edges occur within the same block and are measured by the information against conditional independence conditioned on the rest of the variables in that block and all variables in the previous blocks. Suppose $i \in b_\tau$ so that τ is the index of the block containing i. If j is also in that block, $j \in b_\tau$, the edge is undirected and

$$w_{ij} = \inf(Y_i \perp\!\!\!\perp Y_j | Y_{V(\tau)\setminus\{i,j\}}), \quad \text{where } V(\tau) = \cup_{r\leq\tau} b_r.$$

If j is in a preceding block, the edge is directed.

3 BCS70

The Centre for Longitudinal Studies (2009) gives a short description of the 1970 British Cohort Study (BCS70), some of which we reproduce here. The study is a continuing, multi-disciplinary longitudinal study which takes as its subjects all those living in England, Scotland and Wales. Data were collected on the births and families of just under 17,200 babies. who were born in one particular week in April 1970. Since 1970 there have been six attempts to gather information from the whole cohort: 1975, 1980, 1986, 1996, 2000 and 2004.

With each successive attempt, the scope of enquiry has broadened from a strictly medical focus at birth, to encompass physical and educational development at the age of five, physical, educational and social development at the ages of ten and sixteen, and then to include economic development and other wider factors at 26, 29 and 34 years.

Data have been collected from a number of different sources, and in a variety of ways. In the birth survey, information was collected by means of a questionnaire that was completed by the midwife present at the birth, and supplementary information was obtained from clinical records. The five-year and ten-year surveys were carried out by the Department of Child Health, Bristol University and the survey at these times was named the Child Health and Education Study (CHES) . In 1975 and 1980, parents of the cohort members were interviewed by Health Visitors, and information was gathered from head and class teachers (who completed questionnaires), the school health service (which carried out medical examinations on each child), and the subjects themselves (who undertook tests of ability). In both 1975 and 1980, the cohort was augmented by the addition of immigrants to Britain who were born in the target week in 1970.

Variables of interest

We examine a subset of 2457 individuals which had recorded information on most tests for language and mathematics over the six follow up years. The scores in each year have different scales as the testing procedures changed with different ages. To make these variables comparable and discrete we replace each score by an approximate percentile categorisation. For both language and mathematics and for each year, we classify the score into one of the three percentile bands $(0, 1/3)$, $(1/3, 2/3)$, $(2/3, 1)$ of that score, and label the result 0, 1, 2 respectively. As the scores are integers these percentiles only approximately hold 1/3 of the cases. The missing values are recorded in a fourth category, labelled 3. While this categorisation leads to some suppression of information, much is retained. This gives 12 categorical variables 1975:mat1,lan1, 1980:mat2,lan2, 1986:mat3,lan3, 1996:mat4,lan4, 2000:mat5,lan5, and 2004:mat6,lan6. Two other variables included are sex, and social class at birth (scbirth) recorded in 4 levels.

For example a tabulation of the language percentiles in 1980 and 1986 is shown on the left hand side of Table 1, and of the language and mathematics percentiles in 1986 on the right. From this table it is seen there are a substantial number of missing values, which is typical for many longitudinal studies. In Table 2 we give the number of missing values for language and mathematics.

Table 1: Cross tabulation of the percentiles for language 1980 and 1986 (left) and for mathematics 1986 and language 1986 (right).

	lan3					lan3			
lan2	0	1	2	3	mat3	0	1	2	3
0	155	98	64	127	0	262	179	118	87
1	108	146	151	205	1	153	166	186	136
2	26	52	82	103	2	98	153	194	157
3	302	234	218	386	3	78	32	17	441

Table 2: Missing values for language and mathematics.

lang1	lang2	lang3	lang4	lang5	lang6
45	1140	821	717	1704	2284
math1	math2	math3	math4	math5	math6
69	1182	568	716	1970	2333

Dwigs for language and mathematics

We are interested in the longitudinal dependency structure for language and mathematics considered separately and together, and where the presence of missing values is taken into account. A chain that follows the time line of the cohort is clearly meaningful, so the variables are placed in blocks by year of observation. The initial block contains just sex and social class which are determined at time of birth. We use the code discussed in Whittaker and Kao (2009). Here the divergences are approximated by a deviance calculation from fitting main effect binary logistic regressions.

The longitudinal dependency structure for language and for mathematics is given in Figure 1. Two covariates, sex and social class at birth, form the first block of the chain graph. The lack of an edge between these covariates indicate their marginal independence. The percentile measure lan1 weakly depends on both these covariates, while lan2 is approximately independent of these covariates give lan2. However lan3 depends on social class having adjusted for lan1 and lan2. The percentile lan4 also depends on social class and its preceding language percentiles, but here the dependency from lan3 is rather larger than other divergences. A similar story is associated with lan5 but there is an additional dependency to sex. However lan6 shows a different picture with almost no dependency on previous language percentiles.

The pattern for the mathematics percentiles has a remarkably similar configuration as that for language, with slightly weaker dependences.

Missing values

A worry is that because the missing value indicator is included with the subject percentile, it is the missing values that dictate, or at least modify, the dwigs

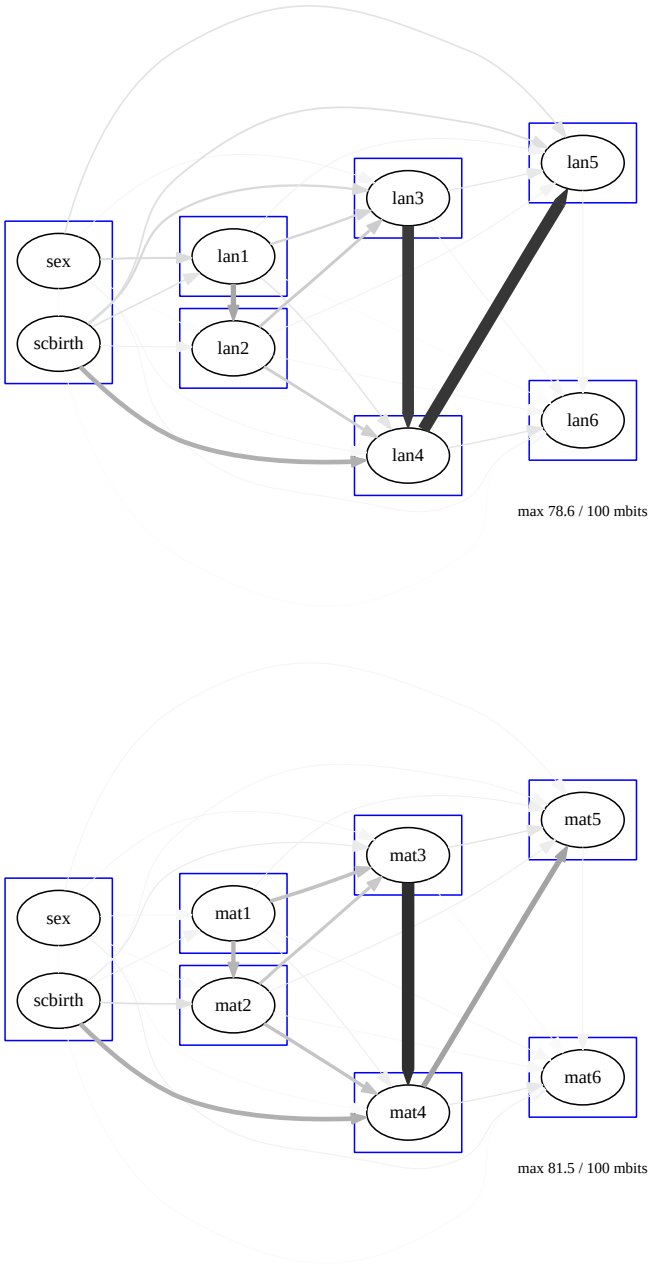


Figure 1: Dwigs exhibiting the longitudinal dependency structure for language (upper panel) and mathematics (lower panel).

above. In principle it is possible to do a complete case analysis but as can be seen from Table 2, there are few complete cases that cover six subject variables.

We make two proposals. The first is to construct the dwig for the longitudinal dependency structure of the missing values. The second is to condition on observing a subject variable, and to examine any dependence on a previous missing value. This differs from a complete case analysis because only a single variable is taken, rather than all six, and one may expect extra power by working with its marginal distribution. While it is impossible to assess whether a measured language or mathematics percentile is associated with its missing value indicator, it is possible to do this for preceding indicators.

The dwig for the longitudinal dependency structure of the missing values is displayed in Figure 2. The indicator is a combined indicator for language and mathematics that takes the value 0 when both subjects are measured, and 1 otherwise. It would be possible to build separate dwigs for separate indicators but there is almost no difference for this data set, and the simplification helps. The graph shows that the missing values in the different years are approximately

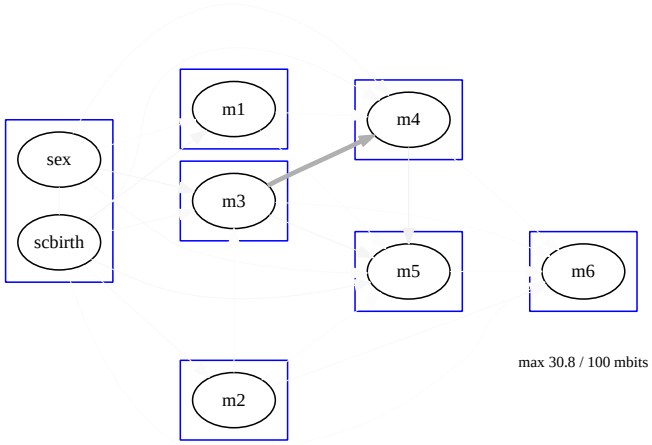


Figure 2: A dwig for the longitudinal dependency structure of the missing values.

independent, with one exception of the transition from 1986 to 1996 (m3 to m4).

An example of the second proposal is displayed in Figure 3, which considers the conditional distribution of the language percentile in 1996 (lan4) given that it is observed, and the preceding missing value indicators. The conditioning is handled by confining the analysis to the subset of data for which lan4<3. No large divergences between the indicators and lan4 are seen in this graph, supporting the hypothesis of missing values occurring at random. Further analyses taking a single language variable and its preceding missing values, and similarly for the mathematics variables, all reached the same independence conclusion.

In this subset of data the divergence between lan4 and social class is larger than in the upper panel of Figure 1. This may be explained because this graph marginalises over lan3 which has a strong effect on lan4 and is also indirectly dependent on social class.

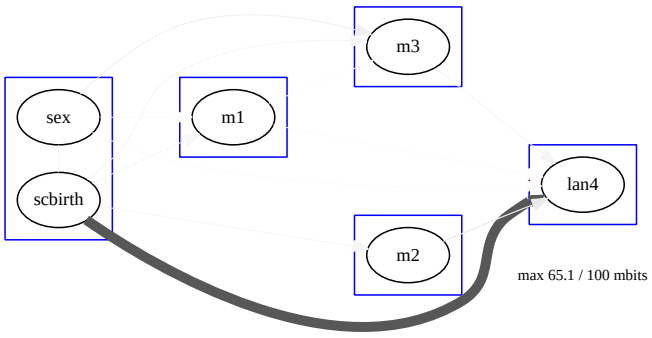


Figure 3: A dwig for the conditional distribution of the language percentile in 1996 (lan4) given that it is observed, and the preceding missing value indicators.

A combined dwig

Somewhat speculatively we give a combined dwig for both language and mathematics in Figure 4. The reason for the hesitancy is that with 14 variables in the graph the conditional distribution of the later dependencies is rather sparse, and hence subject to larger sampling fluctuations. Having said this we note the Figure shows the strong association between language and mathematics in any given year, and that this association is stronger than any connections to previous years. The approximate independence of the variables in the final year of the study is consistent with Figure 1. There may be an argument that the serial dependence in the sequence of the language percentiles is stronger and more linear than that of the mathematics percentiles. Interestingly, but again speculatively, is the assymmetry in the transitions from 1986 (lan3,mat3) to 1996 (lan4,mat4). The transition seems to be between lan3 and lan4, and this might be argued to 'explain' the stronger dependence between mat3 and mat4 visible in Figure 1.

4 Summary

We have shown by example that dwigs give an overview of variables taken from survey data. In the context of the 1970 British Cohort Study we have displayed the parallel dependence structure of mathematics and language percentiles. We have indicated that missing values have little effect on these relationships, though because of their very nature it is impossible to demonstrate that missing values are entirely unrelated to the counter factual value of the unobserved variable. We have demonstrated that the observed variables of interest are unrelated to other missing value indicators, which substantiates part of the assumption that data is missing at random in BCS70. Such an analysis may well generalise to other longitudinal studies.

Acknowledgements: Thanks go to Ian Plewis for introducing the author to the British Cohort Study.

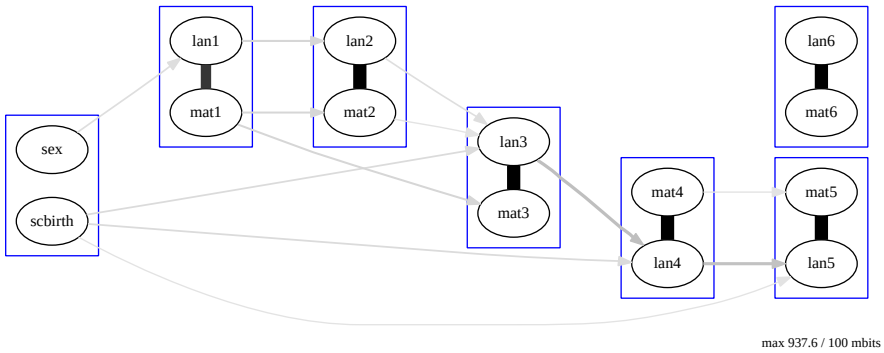


Figure 4: A combined dwig exhibiting the longitudinal dependency structure for the joint distribution of language and mathematics.

References

Centre for Longitudinal Studies, . (2009). *British Cohort Study*. Institute of Education, London, `http://www.cls.ioe.ac.uk` (click on British Cohort Study).

Cover, T. and Thomas, J. (2002). *Elements of Information Theory*. John Wiley and Sons, 2nd edition.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.*, 8:522–539.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, Oxford.

Wermuth, N. and Cox, D. (1996). *Multivariate Dependencies*. Chapman Hall, London.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Whittaker, J. and Kao, C.-F. (2009). Divergence weighted independence graphs: an exploratory tool for survey analysis within a design based framework. *Submitted*.

USING MIXTURES OF TRUNCATED EXPONENTIALS FOR SOLVING STOCHASTIC PERT NETWORKS

Esma Nur Cinicioglu
Istanbul University, Faculty of Business Administration, Quantitative Methods Department, Istanbul, Turkey
esmanurc@istanbul.edu.tr

Prakash P. Shenoy
University of Kansas School of Business, 1300 Sunnyside Ave, Summerfield Hall, Lawrence, KS 66045-7585
pshenoy@ku.edu

Abstract

In this paper, we transform a PERT network into a mixtures of truncated exponentials Bayesian network. We use the Shenoy-Shafer architecture to propagate the MTE potentials in the resulting MTE PERT Bayes net and thus to find the marginal distribution of the project completion time. Finding the distribution of the project completion time is important because there is no closed form expression for the distribution of the maximum of two normal distributions and this fact, previously forced the researchers to make false assumptions about its distribution. In this research, we show that by approximating the maximum of two distributions using MTE's a very accurate estimation for the project completion time can be obtained.

1 Introduction

Large projects contain a series of activities that possess precedence constraints which makes project completion time difficult to manage. One of the most famous project management techniques is *Program Evaluation and Review Technique* (PERT). PERT was invented in 1958 for the POLARIS missile program by the Program Evaluation branch of the Special Projects Office of the U. S. Navy [Malcolm *et al*. 1959]. PERT networks are directed acyclic networks where the nodes represent duration of activities and the arcs represent precedence constraints. The easy applicability of PERT networks to all kind of projects made it widely used in practice. However, although a project may be represented with good accuracy using PERT networks, the accurate estimation of the project completion time is not an easy task to fulfill.

The classical solution [Malcolm *et al*., 1959] for PERT networks assumes that all activities are independent random variables, having approximate beta distributions parameterized by three parameters: mean time m, minimum (optimistic) completion time a, and maximum (pessimistic) completion time b. Using the expected duration times we compute the path that takes the longest time to finish (the critical path), hence the project completion time.

In order to involve uncertainty in the computation of project completion time and hence to improve the accuracy of the estimations, Sculli [1983] suggested to assume that all activity durations are independent, having the

Gaussian distribution. This suggestion is good in the sense that it involves the uncertainty of activity durations in the computation of the project completion time. However, with this method it is also assumed that the distributions of the activity completion times are Gaussian. The completion time of an activity i is given by $C_i = \text{Max}\{C_j \mid j \in \Pi(i)\} + D_i$, where C_j denotes the completion time of activity j, D_j denotes the duration of activity j, and $\Pi(i)$ denotes the parents (immediate predecessors) of activity i. The maximum of two independent Gaussian random variables is not Gaussian, but the distribution of C_i is assumed to be Gaussian with the parameters estimated from the parameters of the parent activities. The current methods in the literature fail to recognize the true distribution of the maximum of two independent distributions and thus make false assumptions, like the maximum of two normal distributions are again normally distributed. Depending on the value of parameters this assumption can lead to large errors for the completion time of the activities which will lead to inaccurate estimates for the project completion time.

Motivated by this problem in the literature, Cinicioglu and Shenoy [2006] provided a new method which aims to approximate the true distribution of the project completion time by eliminating the false assumptions for the distribution of the maximum of two Gaussians. With this method, a PERT network is transformed into a mixtures of Gaussians Bayesian network and then Lauritzen-Jensen algorithm is used to make inferences in the resulting MoG Bayesian network. Mixtures of Gaussians (MoG) hybrid Bayesian networks [Lauritzen, 1992] are Bayesian networks with a mix of discrete and continuous variables. In MoG Bayesian networks the discrete variables cannot have continuous parents, and all continuous variables have the so-called conditional linear Gaussian distributions.

Representation of a PERT network as a MoG Bayesian network is beneficial in the sense that it eliminates the false assumption made in the literature which assumes that the maximum of two normally distributed independent random variables is again normally distributed. However, the transformation process of a PERT network into a MoG Bayesian network is cumbersome because of the restricted nature of MoG Bayesian networks. The inability of discrete variables to have continuous parents and the enforcement for continuous variables to possess conditional linear Gaussian distributions makes the transformation process of a PERT network into a MoG Bayes net too complex for practical use.

For that reason, in this research we work on a different method, an alternative to MoG Bayesian networks, which overcomes the difficulties involved in solving stochastic PERT networks using MoG's, but still possess the advantages involved in it. The alternative we suggest in this paper for solving stochastic PERT networks with MoGs, is to solve them using mixtures of truncated exponentials (MTE). We proceed as follows: First we transform a PERT network into a PERT Bayes net, so we can model the dependencies between activity durations. Next, we transform the PERT Bayes net into a MTE network by approximating the activity durations using MTE's. Finally using the Shenoy-Shafer architecture we propagate the MTE potentials and find the marginal distribution of the project completion time. To evaluate our method we

compare the mean and variance of the marginal distribution of the project completion time with the exact analytic results using Clark's method [1961] and the shape of our distribution with the actual distribution calculated by brute force using order statistics.

2 Representation of a PERT network as a Bayesian network

In order to demonstrate our method of solving stochastic PERT networks using mixtures of truncated exponentials we will use a simple example of a PERT network and compute the marginal distribution of the project completion time. Consider the PERT network given in Figure 1 below. This network represents a project with the activities A_1, A_2 and A_3. S stands for the project start time and E stands for the project completion time. We assume that the project start time is zero. The precedence constraints, represented by arcs, are as follows: The activities A_1 and A_2 do not have any predecessors. The activity A_3 can only be started after A_1 is completed.

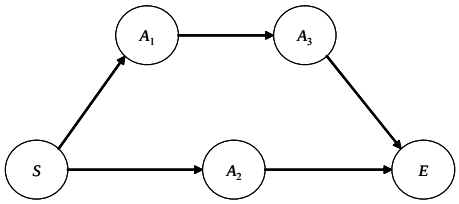


Figure 1. An example of a stochastic PERT network with three activities

The distributions of activity durations are known, and we are informed that the activity durations A_1 and A_3 are positively correlated. Following the method described in Jenzarli[1995] this PERT network will be transformed into a PERT Bayesian network in four basic steps, allowing us to model the dependencies between the activity durations.

Let D_i and C_i denote the duration and the completion time of the activity i, respectively. As the first step of the transformation process, the activity durations are replaced with activity completion times. Next, activity durations will be added with an arrow from D_i to C_i, so that each activity will be represented by two nodes, its duration D_i and its completion time C_i. As the next step, notice that the completion times of the activities which do not have any predecessors will be the same as their durations. Hence, these activities A_1 and A_2 will be represented just by their durations, as D_1 and D_2. Remember that we are informed that the activities D_1 and D_3 are positively correlated. As the last step of the transformation process, the dependency between these activity durations will be depicted by adding an arrow from D_1 to D_3. We assume that the project start time is zero with probability 1 and each activity will be started as soon as all the preceding activities are completed. Accordingly, E represents the completion time of the project, which is the $Max\{D_2, C_3\}$. The resulting

PERT Bayes net is given in Figure 2 below. Notice that the deterministic variables, C_3 and E, are depicted as double bordered ovals. The next section describes mixtures of truncated exponentials.

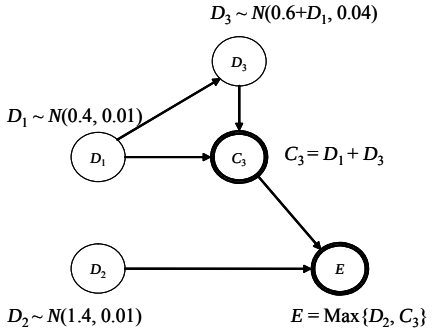


Figure 2: An example of a PERT Bayesian network

3 Mixtures of Truncated Exponentials

MTE's are an alternative to discretization and Monte Carlo methods for solving hybrid Bayesian networks [Moral *et al.* , 2001; Rumi, 2003]. MTE potentials can be used for inference in hybrid Bayesian networks that do not fit the restrictive assumptions of the conditional linear Gaussian (CLG) model, such as networks containing discrete nodes with continuous parents.

A mixture of truncated exponential (MTE) [Moral *et al.* , 2001; Rumi, 2003] has the following definition.

Let X be a mixed n-dimensional random variable. Let $Y = (Y_1, \ldots, Y_d)$ and $Z = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of X, respectively, with $c + d = n$. A function ϕ: Ω_X α R^+ is an MTE potential if one of the next two conditions holds:

The potential ϕ can be written as

$$\phi(x) = \phi(y, z) = a_0^y + \sum_{i=1}^{m} a_i^y \exp\left(\sum_{j=1}^{c} b_j^y z_j\right) \qquad (3.1)$$

where a^y_0, a^y_i and b^y_j are real numbers for all $i = 1,\ldots,m$, $j = 1,\ldots, c$, $y \in \Omega_Y$ and $z \in \Omega_Z$.

There is a partition $\Omega_1, \ldots, \Omega_k$ of Ω_X verifying that the domain of continuous variables, Ω_Z, is divided into hypercubes, the domain of the discrete variables, Ω_Y, is divided into arbitrary sets, and such that ϕ is defined as $\phi(x) = \phi_i(x)$ if $x \in \Omega_i$, where each ϕ_i, $i = 1, \ldots, k$ can be written in the form of equation (3.1)

In the definition above, k is the number of pieces and m is the number of exponential terms in each piece of the MTE potential.

The nice thing about MTE's is that any probability density function can be approximated by an MTE potential, which can always be marginalized in closed

form. Consider a normally distributed random variable X with mean μ and variance $\sigma^2 > 0$. The PDF for the normal distribution is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-1/2\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

A general formulation for a 2-piece, 3-term unnormalized MTE potential which approximates the normal PDF is as follows [Cobb and Shenoy, 2006a].

$$\psi'(x) = \begin{cases} \sigma^{-1}(-0.010564 + 197.055720\exp\{2.2568434(\frac{x-\mu}{\sigma})\} \\ -461.439251\exp\{2.3434117(\frac{x-\mu}{\sigma})\} \\ +264.793037\exp\{2.4043270(\frac{x-\mu}{\sigma})\}) & \text{if} \quad \mu - 3\sigma \le x < \mu \\ \sigma^{-1}(-0.010564 + 197.055720\exp\{-2.2568434(\frac{x-\mu}{\sigma})\} \\ -461.439251\exp\{-2.3434117(\frac{x-\mu}{\sigma})\} \\ +264.793037\exp\{-2.4043270(\frac{x-\mu}{\sigma})\}) & \text{if} \quad \mu - 3\sigma \le x < \mu \\ 0 \quad \textit{otherwise} \end{cases} \tag{3.2}$$

In the following sections the PERT network example will be transformed into a MTE PERT Bayesian network and solved using the Shenoy-Shafer architecture. The operations necessary to carry out propagation in MTE networks using the Shenoy-Shafer architecture are described in the following, subsection 3.1.

3.1 Operations in MTE Networks

This section describes the operations of restriction, combination, marginalization, normalization, operations with linear deterministic equations and finding the maximum of two distributions using MTE's. These operations are necessary to carry out propagation in our MTE network example. The class of MTE potentials is closed under these operations which allows us to use the Shenoy-Shafer architecture [Shenoy and Shafer, 1990] to propagate the MTE potentials in the network. The definitions of restriction, combination, marginalization and normalization are described in Moral *et al*. [2001]. The operations with linear deterministic variables in MTE networks are described in Cobb and Shenoy[2005]. The operations for finding the maximum of two distributions using MTE's are first described here.

3.1.1 Restriction

Restriction is the operation of entering evidence during the propagation. In restriction, known variables are substituted with their values.

Let ϕ be an MTE potential for $X = Y \cup Z$. Suppose we receive the evidence for a set of variables $X' = Y' \cup Z' \subseteq X$, s.t. its values $x^{\downarrow\Omega}x'$ are as follows: $x' = (y', z')$. After receiving the evidence the values of the variables are known. Accordingly, the potential ϕ should be updated. The new potential defined on $\Omega_{X\setminus X'}$ is as follows:

$$\phi^{R(X'=x')}(w) = \phi^{R(Y'=y',\, Z'=z')}(w) = \phi(x) \qquad (3.3)$$

for all $w \in \Omega_{X\setminus X'}$ such that $x \in \Omega_X$, $x^{\downarrow\Omega}X\setminus X' = w$ and $x^{\downarrow\Omega}X' = x'$. In this definition each occurrence of X' in ϕ is replaced with x'. An example for restriction is provided in section 6.

3.1.2 Combination

MTE potentials are combined by pointwise multiplication. Let ϕ_1 and ϕ_2 be the MTE potentials for $X_1 = Y_1 \cup Z_1$ and $X_2 = Y_2 \cup Z_2$. The combination of ϕ_1 and ϕ_2 is a new MTE potential for $X = X_1 \cup X_2$ defined as follows:

$$\phi(x) = \phi_1(x^{\downarrow X_1})\, \phi_2(x^{\downarrow X_2}) \text{ for all } x \in \Omega_x \qquad (3.4)$$

3.1.3 Marginalization

MTE potentials are marginalized by summing over discrete variables and integrating over continuous variables. Let ϕ be an MTE potential for $X = Y \cup Z$. The MTE potentials are closed under marginalization, so the marginal of ϕ for the set of variables $X' = Y' \cup Z' \subseteq X$ is a MTE potential which is computed as follows:

$$\phi^{\downarrow X'}(y', z') = \sum_{y \in \Omega_{Y\setminus Y'}} \left(\int_{\Omega_{Z\setminus Z'}} \phi(y, z)\, dz'' \right) \qquad (3.5)$$

where $z = (z', z'')$, and $(y', z') \in \Omega_{X'}$. The variables can be marginalized in any sequence, discrete before continuous or continuous before discrete as shown in Formula 3.5.

In the process of marginalization, when the limits of integration include linear functions, then we may end up with linear terms in the remaining variables. These linear terms can be replaced with an MTE approximation so that the result of the marginalization is again an MTE potential. For a linear term x defined over the domain $[x_{min}, x_{max}]$, we replace x with

$$x_{\min} + (x_{\max} - x_{\min})(0.5 * (-13.5070292 + 13.5070292\, Exp[\frac{0.0726981(x - x_{\min})}{(x_{\max} - x_{\min})}] \qquad (3.6)$$
$$+ 0.5 * (13.5070364 - 13.5070364 Exp[\frac{-(0.0754406(x - x_{\min})}{(x_{\max} - x_{\min})}]$$

The replacement of the linear terms ensures that MTE potentials are closed under marginalization.

3.1.4 Normalization

Let $X = Y \cup Z$ be a set of variables where Y is a discrete and Z is a continuous variable. Let ϕ' be the MTE potential for X. Normalization constant for K is calculated as follows:

$$K = \sum_{y \in \Omega_Y} \left(\int_{\Omega_Z} \phi'(y, z) dz \right) \qquad (3.7)$$

If join trees are initialized with normalized potentials the normalization constant equals to one when no evidence is observed.

3.1.5 Linear Deterministic Equations

If the variable being deleted is contained in a linear deterministic equation in the network, then the marginalization operation is different. If it is the case, then we solve the equation for the variable being deleted and then substitute this solution in the updated potentials in the network.

Let ψ denote the distribution of $Y|x \sim f_{Y|X}$ and let ζ denote the equation $Z = X + Y$. Suppose we want to delete the variable Y from the network. By solving the equation for Y and substituting the solution in $f_{Y|x}$ we can remove Y out of the combination and hence find the distribution of $Z|x$. The details are as follows:

$(\zeta \otimes \psi)^{-Y} = ([Z = X + Y] \otimes f_{Y|X}(y))^{-Y} = ([Y = Z - X] \otimes f_{Y|X}(y))^{-Y} = f_{Y|x}(z - x)$

3.1.6 Maximum of Two Distributions

Finding the distribution of the maximum of two or more distributions has been the interest of many communities of researchers. Especially in the domains of project management, this problem occupies an important place since the completion time of an activity is the sum of its duration and the maximum between the completion times of its immediate predecessors. For this reason, it can be concluded that an accurate estimation of the project completion time is very much affected by an accurate estimation of the activity completion times.

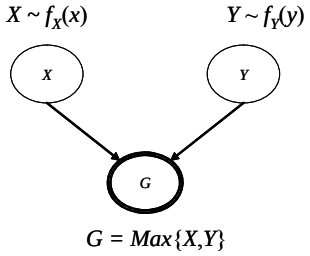


Figure 3. Maximum of two distributions

The marginal probability density function of the maximum of two distributions can be computed by brute force using order statistics. Consider the small BN given in Figure 3. X and Y are continuous variables which have density functions $f_X(x)$ and $f_Y(y)$, respectively. G is a deterministic variable which is distributed as $G = Max\{X, Y\}$. Let F_G denote the cumulative distribution function (CDF) of G, F_X denote the CDF of X and F_Y denote the CDF of Y. Then, $F_G(g) = F_X(g)F_Y(g)$. Therefore, the probability density function of G is given by $f_G(g) = (d/dg)F_G(g) = f_X(g)\, F_Y(g) + F_X(g)\, f_Y(g)$, where f_X and f_Y are the PDFs of X and Y, respectively. Since there is no closed form expression

for the CDF of a normal distribution, there is no closed form expression for $f_G(g)$ when X and Y are normally distributed. Since MTE potentials are closed under integration both $F_X(g)$ and $F_Y(g)$ can be expressed as MTE potentials. And since MTE potentials are closed under multiplication and addition $f_G(g)$ can also be expressed as MTE potentials. Then, by using the MTE approximations of X and Y, we can obtain an MTE approximation for the distribution of $f_G(g)$.

The next section describes the transformation of our PERT Bayes net example into a MTE PERT Bayesian network.

4 Transformation of a PERT Bayesian network into a MTE PERT Bayesian network

The primary objective of this study is to compute the completion time of the project without setting any assumptions for activity distributions. This objective will be materialized by approximating the activity durations using mixtures of truncated exponentials and propagating the resulting mixtures of truncated exponentials network using the Shenoy-Shafer architecture.

Consider the PERT Bayes net given in Figure 2. Notice that it is not a MTE Bayesian network since the activity durations D_1, D_2, and D_3 are all normally distributed. In order to transform this PERT Bayes net into a MTE Bayesian network all of these activities will be approximated using MTE's. The MTE approximation of D_1 overlaid on the actual normal distribution is given in Figure 4 below.

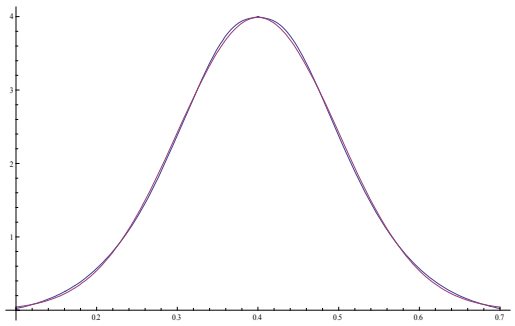


Figure 4. The actual distribution of D_1 overlaid on its MTE approximation

The probability distribution for D_3 is defined as $D_3|d_1 \sim N(0.6+d_1, 0.04)$. The plot for the MTE approximation for D_3 is given in Figure 5 below.

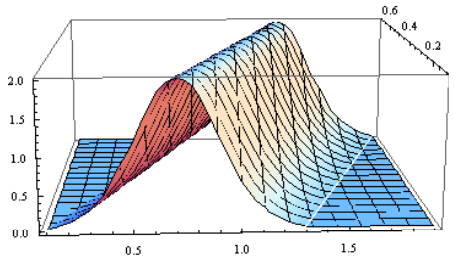


Figure 5. MTE approximation for $D_3|d_1$

5 Fusion Algorithm

The fusion algorithm, first described by Cannings *et al.* [1978], is used to compute the marginal for a variable using local computation [Shenoy, 1992]. Shenoy [1997] described the fusion algorithm as a guide to construct join trees where Shenoy-Shafer architecture will be used to compute the marginals of the variables. The basic idea of the fusion algorithm is to delete all the variables in the network successively, until we end up with the marginal distribution of the variable of interest.

In this research, we are interested in computing the marginal distribution of the project completion time. Hence, using fusion algorithm, the variables in the MTE PERT Bayes net will be deleted successively, until we end up with the marginal distribution of the project completion time, F. Though different deletion sequences may lead to different computational efforts, the outcome of the network does not get affected with the deletion sequence used. In this example, we will use the deletion sequence D_3, D_1, (D_2, C_3) in order to find the marginal distribution of the project completion time. Figure 6 illustrates the construction of the join tree for the PERT example.

The details of the messages necessary to compute the marginal distribution of the project completion time are as follows:

Fusion with respect to D_3:

Fusion w.r.t. D_3, refers to removing the variable D_3 from the network. This will be done first by combining all the potentials that contain D_3 and next by removing D_3 out of the combination by marginalizing the combination down to the remaining variables. Let $f_{D_3|d_1}$ denote the distribution of $D_3|d_1$. Let χ_3 denote the equation for $C_3 = D_1 + D_3$. By solving the equation for D_3 and substituting D_3 in $f_{D_3|d_1}$ we can find the distribution of $C_3|d_1$. The details are as follows:

$C_3 = D_1 + D_3$

$D_3 = C_3 - D_1$

$f_{C_3|d_1}(c_3) = f_{D_3|d_1}(c_3 - d_1)$

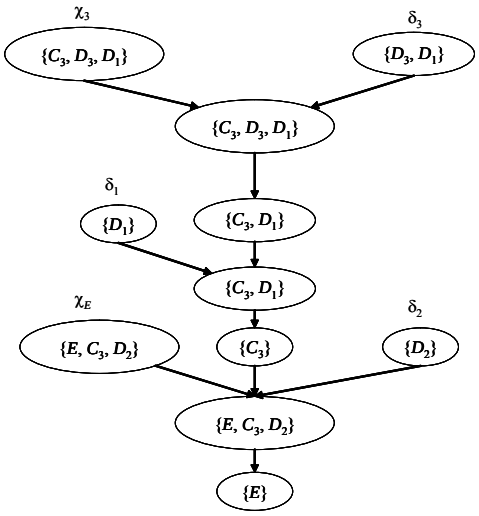


Figure 6. Creation of the binary join tree using the fusion algorithm.

Fusion with respect to D_1:

The variables whose domains contain D_1, (D_1 itself and $C_3|d_1$), are both continuous variables, so deleting D_1 from the network involves finding the joint $f_{C_3,\ D_1}(c_3,\ d_1)$ and integrating this combination over the domain of D_1. The details are as follows:

$$f_{C_3,\ D_1}(c_3,\ d_1) = f_{C_3|d_1}(c_3)\ f_{D_1}(d_1)$$

$$(f_{C_3,\ D_1}(c_3,\ d_1))^{\downarrow C_3} = \int f_{C_3,D_1}(c_3,\ d_1)\ dd_1 = f_{C_3}(c_3)$$

The expected value and variance for the marginal of C_3 are calculated as 1.4 and 0.0786. These answers are comparable with results from multivariate normal theory, which gives an expected value and variance of 1.4 and 0.08.

The next step is to find the marginal distribution of $E = Max\{C_3,\ D_2\}$ which requires the variables, C_3 and D_2, to be deleted at the same time.

Figure 7 represents the current state of our network after the variables D_3 and D_1 are removed from the network. As the next and final step, we have to find the project completion time $E = Max\{C_3,\ D_2\}$ which requires the variables C_3 and D_2 to be deleted at the same time.

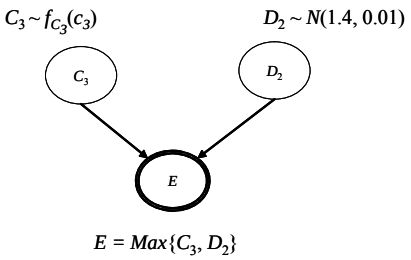


Figure 7. The conditional distribution of E after D_3 and D_1 are deleted from the network

As explained in subsection 3.1.6, the probability density function of F_E is given by $f_E(e) = (d/de)F_E(e) = f_{C_3}(e)\, F_{D_2}(e) + F_{C_3}(e)\, f_{D_2}(e)$, where f_{C_3} and f_{D_2} are the PDFs of C_3 and D_2, respectively. In sections 4 and 5 the PDF's of D_2 and C_3 are approximated using MTE's. As the next step of our analysis, we calculate the CDF's of both D_2 and C_3 which we later use for the calculation of the marginal distribution of the project completion time, $f_E(e)$. The plot of the MTE approximation for the CDF of D_2 is illustrated in Figure 8 below.

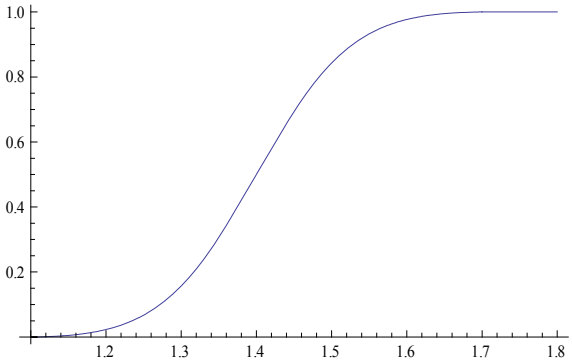


Figure 8. MTE Approximation for $F_{D_2}(e)$

The MTE approximation of $f_E(e)$ overlaid on the actual distribution is given in Figure 9 below.

By comparing the means and variances of the approximation with the exact analytic results calculated with Clark's method [1961], we can evaluate the goodness of our approximation for the marginal distribution of the project completion time, $f_E(e)$. Accordingly, using our method described in this paper the mean and the variance of the marginal distribution of E is calculated as 1.51883 and 0.0300638, respectively. Comparing it to 1.51968 and 0.0306761 given by the exact analytic results, the approximation can be considered as quite successful.

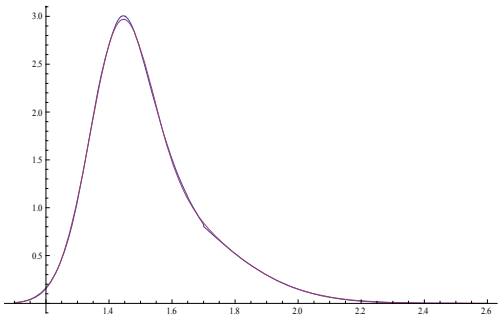


Figure 9. Approximation of $f_E(e)$overlaid on the actual distribution

After normalization, when the limits of integration include linear terms, then we may end up with linear terms in the remaining variables as it is the case with the approximation of C_3 and of the CDF of D_2. These linear terms can be approximated again using MTE potentials, which ensures that the result is again an MTE approximation and MTE's are closed under marginalization. However, replacing the linear terms with the MTE potentials causes bad accuracy in our approximations.

6 Entering Evidence in a MTE PERT Network

In this research MTE PERT Bayes nets are described as an alternative method to solve stochastic PERT networks with which we can compute the marginal distribution of the project completion time without setting any false assumptions for the activity completion times. In this context, it is natural to question our methods described in this research and ask for the advantage obtained by using the methods described, instead of using straight forward simulation methods that are already handy.

With simulation methods the activity durations can be represented realistically. As it is the case with our methods, the activity durations can have any type of distribution and one can also represent the correlation between the activity durations. However, with straight-forward Monte Carlo simulation methods we can not include the observations of continuous variables and update our inferences accordingly. By transforming the PERT network into a MTE Bayesian network and solving it using the Shenoy-Shafer architecture we can update our network, once evidence is observed, and find the posterior distributions of the activities which in turn will result in more accurate estimates for the project completion time.

Consider the PERT Bayesian network given in Figure 10. This is a PERT Bayes net with four activities A_1, A_2, A_3 and A_4. Notice that the activities are depicted by their durations, as D. Suppose we know that the activities A_1 and A_2 will be performed by the same contractor. The quality of the work done by this contractor is distributed as $f_Q(q)$. The quality of the work performed by the contractor effects the duration of the activities A_1 and A_2 such that with higher quality it will take less time to complete these activities. In addition to these, we also have the information that the same contractor performs another activity

similar to ours within the firm. This activity A_4 is outside of our project but we included it in our network in Figure 10 anyway since it will effect our later conclusions. As you can see in Figure 10 the duration of activity A_4 also depends on the quality of the contractor's job.

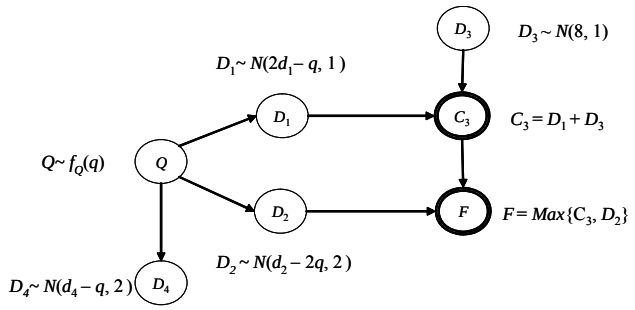


Figure 10. Representation of the example as a PERT Bayesian network

The example described above can be solved using the means of simulation methods as well as with the methods represented throughout this research. However, suppose we observe that the duration of activity A_4 lasted 10 days to complete. Hence we have the evidence $e_{D_4} = 10$. With the methods described in this dissertation this evidence can be incorporated in the network and the estimates for the durations can be updated accordingly, which is not possible using the straight forward simulation. With our method we can find the posterior distribution of Q after receiving the evidence e_{D_4} which in turn will change the estimates for the distributions of A_1 and A_2 and consequently the estimate for the project completion time. Including the observations in the network and updating the distributions accordingly will improve the quality of the inference. The PERT BN after receiving the evidence e_{D_4} is represented in Figure 11 below.

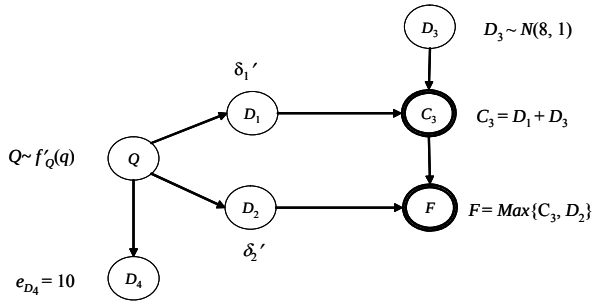


Figure 11. The PERT Bayesian network after receiving the evidence e_{D_4}

7 Summary and Conclusions

Mixtures of truncated exponentials are an alternative tool to mixtures of Gaussians (MoG) to make inferences in stochastic PERT networks. Both MoG's and also MTE's are able to find accurate estimations for the maximum of two

distributions and hence for the project completion time. However, the inference process using MTE PERT networks, compared to MoG's, is much more straightforward in the sense that the MTE PERT networks do not force restrictive settings like, the inability of discrete variables to have continuous parents as it is the case with MoG networks. This fact makes the use MTE PERT Bayes nets better suited for practical use.

Comparing our method to straight forward simulation on the other hand, the MTE PERT Bayesian networks possess the advantage that the observations can be integrated to the inference process. Once evidence is observed we can update our network accordingly and find the posterior distributions of the activities and thus obtain a more accurate estimation for the project completion time.

The drawback with our method is on the other hand, that the number of exponential terms increases rapidly as the fusion algorithm is applied which in turn makes the inference process more difficult to apply. Additionally, in the process of marginalization, when the limits of integration include linear functions, we may end up with linear terms in the remaining variables. These linear terms can be approximated using an MTE approximation and it can be ensured that the result is again an MTE potential. However, replacing the linear terms with the MTE potentials causes bad accuracy in our approximations.

References

[1] Cannings,C., E. A. Thompson and M.H. Skolnick (1978), "Probability functions on complex pedigrees", *Advances in Applied Probability* , 10, 26-61

[2] Cinicioglu. E.N., Shenoy, P.P (2006), "Solving Stochastic PERT Networks Exactly Using Hybrid Bayesian Networks," in J. Vejnarova and T. Kroupa (eds.), *Proceedings of the Seventh Workshop on Uncertainty Processing (WUPES-06)*, pp. 183--197, 2006, Mikulov, Czech Republic, Oeconomica Publishers

[3] Clark, C. E. (1961), "The greatest of a finite set of random variables," *Operations Research*, **9**(2), 145–162.

[4] Cobb, B. R. and P. P. Shenoy (2005), "Hybrid Bayesian networks with linear deterministic variables," in F. Bacchus and T. Jaakkola (eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-First Conference* (UAI-05), 136–144, AUAI Press, Corvallis, OR.

[5] Cobb, B. R. and Shenoy, P.P (2006a)"Inference in Hybrid Bayesian Networks with Mixtures of Truncated Exponentials," International Journal of Approximate Reasoning, Vol. 41, No. 3, pp. 257--286,

[6] Jenzarli, A. (1995), "Modeling dependence in project management," PhD dissertation, University of Kansas School of Business, Lawrence, KS.

[7] Lauritzen, S. L. (1992), "Propagation of probabilities, means and variances in mixed graphical association models," *Journal of American Statistical Association*, 87(420), 1098–1108.

[8] Malcolm, D. G., J. H. Roseboom, C. E. Clark, and W. Fazar (1959), "Application of a technique for research and development program evaluation," *Operations Research*, **7**, 646–669.

[9] Moral, S., Rumí, R., Salmeron, A. (2001) "Mixtures of truncated exponentials in hybrid Bayesian networks, Symbolic and Quantitative Approaches to Reasoning Under Uncertainty, Lecture Notes in Artificial Intelligence, Vol.2143, Springer Verlag, Heidelberg, pp. 156-167

[10] Rumí, R. (2003) "Modelos De Redes Bayesianas Con Variables Discretas Y Continuas", Doctoral Thesis, Universidad de Almariá, Departamento de Estadística Y Matemática Aplicada, Almería, Spain

[11] Sculli, D. (1983), "The completion time of PERT networks," *Journal of the Operational Research Society*, 34(2), 155–158.

[12] Shenoy, P. P. and G. Shafer (1990), "Axioms for Bayesian and belief function propagation," in R. D. Shachter, T. S. Levitt, L. N. Kanal and J. F. Lemmer (eds.), *Uncertainty in Artificial Intelligence 4*, 169–198, North-Holland, Amsterdam.

[13] Shenoy, P.P (1992), "*Valuation-based systems: A framework for managing uncertainty in expert systems*", in Zadeh, L. A. and J. Kacprzyk(eds.), fuzzy Logic for the Management of Uncertainty, 83-104, John Wiley &Sons, New York, NY

[14] Shenoy, P.P (1997), "Binary Join Trees for Computing Marginals in the Shenoy-Shafer Architecture", International Journal of Approximate Reasoning, **17**(2-3), 239-263

Title: Proceedings of WUPES 2009

Publisher: University of Economics, Prague

Editors: Tomáš Kroupa, Jiřina Vejnarová

Cover design: Jiří Přibil

ISBN 978-80-245-1543-4