Proceedings of the 10th Workshop on Uncertainty Processing *(WUPES'15)* Monínec, Czech Republic

Václav Kratochvíl
(editor)

September 16-19, 2015

Published by:
Vysoká škola ekonomická v Praze,
Nakladatelství Oeconomica, 2015

Organized and sposored by:
Faculty of Management, University of Economics, Jindřichův Hradec
Institute of Information Theory and Automation, Czech Academy of Sciences

Credits:
Cover design: Jiří Přibil
Editor: Václav Kratochvíl
LaTeX editor: Václav Kratochvíl
using LaTeX's 'confproc' package, version 0.7 (by V. Verfaille)

Printed by Typos, tiskařské závody, s. r. o., Plzeň — September 2015

First edition

ISBN 978-80-245-2102-2

10th WORKSHOP ON UNCERTAINTY PROCESSING

Organized by:

Faculty of Management
University of Economics

and

Institute of Information Theory and Automation
Czech Academy of Sciences

Monínec

September 16-19, 2015

Programme Committee Chair:
Milan Studený, *Institute of Information Theory and Automation, Czech Academy of Sciences*

Seminar Chair:
Radim Jiroušek, *University of Economics, Faculty of Management, Jindřichův Hradec*

Programme Committee:
Nihat Ay *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*
Gernot D. Kleiter *Department of Psychology , University of Salzburg, Austria*
Vilém Novák *Centre of Excellence IInnovations, University of Ostrava, Czech Rep.*
Prakash P. Shenoy *School of Business, The University of Kansas, Lawrence, KS, USA*

Organizing Committee:
Václav Kratochvíl, *Czech Academy of Sciences, Prague*
Iva Krejčová, *University of Economics, Faculty of Management, Jindřichův Hradec*
Tomáš Kroupa, *Czech Academy of Sciences, Prague*
Milan Studený, *Czech Academy of Sciences, Prague*
Lucie Váchová, *University of Economics, Faculty of Management, Jindřichův Hradec*
Jiřina Vejnarová, *Czech Academy of Sciences, Prague*
Jiří Vomlel, *Czech Academy of Sciences, Prague*

Foreword

Writing the opening words to the Proceedings of the 10th Workshop on Uncertainty Processing (WUPES 2015) is a natural opportunity to take a glance back at the 27-year history of these scientific meetings. Though the organization of the first meeting of this series was a reaction of defiance to the fact that we could not attend the (1st) Conference on Uncertainty Processing and Management of Uncertainty held in 1986 in Paris, it finally appeared to be a good idea to start organizing small informal meetings with a limited number of participants. This has enabled us to keep the working atmosphere, which is supported by the printed Proceedings of the preliminary versions of the lectures presented that are distributed to the participants at the beginning of the meeting. There are no parallel sessions and the participants have enough time for personal discussions above and beyond the Proceedings, either during coffee breaks, lunch time, or on the occasion of an excursion, which are integral parts of the WUPES programme. Starting in 1994, the meetings have been organized at different interesting places in the Czech Republic as indicated in the following survey.

1) *1988, June 20-23,* Alšovice
2) *1991, September 9-12,* Alšovice
3) *1994, September 11-15,* Třešť
4) *1997, January 22-25,* Praha
5) *2000, June 21-24,* Jindřichův Hradec
6) *2003, September 9-12*, Hejnice
7) *2006, September 16-20,* Mikulov
8) *2009, September 19-22,* Liblice
9) *2012, September 12-15,* Mariánské Lázně
10) *2015, September 19-22,* Monínec

Thus, we can see the meetings have been held in several historical castles, one monastery, and one of the best-known Czech spa towns. This year, we are meeting in a sports centre, Monínec, the geographic barycentre of Bohemia.

Having data describing ten cases, statisticians would be able to present an extended statistical survey from the history of WUPES meetings. But let us stop looking back in history and turn our attention to the actual meeting. Based on the extended abstract, the Programme Committee accepted 23 papers for presentation at the meeting. Besides traditional topics like the fuzzy-set approach to the coherence theory, conditional independence, compositional models, imprecise probabilities (including belief functions) and the marginal problem, new themes have emerged, namely the applications of

graphical models (in particular Bayesian networks and phylogenetic trees), methods of polyhedral geometry and learning and troubleshooting algorithms.

As usual, the best papers will be selected and their authors will be invited to submit their extended versions for publication in the special issue of a reputable international journal. This year we have made a preliminary agreement with the International Journal of Approximated Reasoning. Nevertheless, reading the contributions from these Proceedings one should keep in mind the working character of the meeting. Let us stress that it is also a tradition that contributions presenting preliminary achievements from on-going projects and not-yet-finished results have been invited. Namely, such contributions spur exciting discussions, and thus we believe that this year's Workshop will be at least as pleasant and successful as those in the previous years.

WUPES 2015 is organized jointly by the Institute of Information Theory and Automation of the Czech Academy of Sciences and by the Faculty of Management, University of Economics, Prague. It is quite natural that such a meeting could not materialize if it were not for the hard work of many our colleagues and friends. This is why we want to express our gratitude to all the members of both the Programme and Organizing Committees. Last but not least, we also want to acknowledge the fact that this workshop is organized, due to the fact that the research of several members of the Organizing Committee is financially supported by grants GA ČR no 15-00215S and 13-20012S.

Radim Jiroušek *Václav Kratochvíl* *Milan Studený*

Papers

Extended Abstracts

Properties of composition for continuous variables

Vladislav Bína
Faculty of Management in Jindřichův Hradec
University of Economics in Prague
Jarošovská 1117/II, 37701 Jindřichův Hradec
e-mail: bina@fm.vse.cz

Abstract

Author presents an operator of composition for densities of continuous random variables and analyzes its properties, i.e., particularly the assertions useful for marginalization, concerning the conditional independence, entropy and alteration of ordering of composed densities in the model. He proposes generalized function of Dirac delta as a degenerated distribution allowing to express operations of conditioning and intervention using composition of Dirac delta with continuous density or compositional model.

Keywords: Operator of composition, continuous random variable, Dirac delta function, causality, conditioning, intervention.

1 Introduction

The paper presents an operator of composition for densities of continuous random variables (defined in [1]) and analyzes its basic properties in a manner analogous to [3]. The basic aim is to introduce well-known generalized function of Dirac delta as a (degenerated) distribution allowing to define operation of conditioning and intervention (for discrete case see [2]) in compositions of continuous densities.

But Dirac delta also provides a possibility to include deterministic variables into the models, to define mixed random variables (partially continuous with steps in probability density function) and to include discrete distributions to compositional models build from densities of continuous random variables. This unifying element thus allows us to approximately model the problems including continuous random variables and has a potential to make such approximations computationally feasible.

2 Preliminaries and basic notions

Throughout the paper we consider finite set of random variables $(X_1, \dots, X_n)$ with values or vectors of values denoted by corresponding lowercase letter. The domain of variables will be denoted by corresponding bold uppercase letter $\mathbf{X}_i$. Variables with finite or countable set of possible *states* are called *discrete*, other variables are called *continuous* (at which we particularly aim). Both discrete and continuous variables can be described using (multi-dimensional) *cumulative distribution function* (CDF). But for definition of operator of composition in discrete case we need *probability mass function* which is in theory of compositional models usually denoted by small Greek letters $(\kappa, \lambda, \mu, \nu, \pi, \dots)$. The theory of discrete compositional models and all related notions can be found in [3].

In order to define the operator of composition of continuous variables we require its cumulative distribution function $F(X)$ to be absolutely continuous and thus must be differentiable almost everywhere. The corresponding *probability density function* is then

$$f(x) = \frac{d}{dx} F(x).$$

Similarly, in case of multivariate (joint) probability density function we require CDF to be sufficiently differentiable and use multiple partial derive.

As we already hinted, the probability density functions (of continuous random variables) will be denoted by small letters of Latin alphabet $(f, g, h, \dots)$. E.g., the abbreviated notation $f(x_K)$ denotes a multidimensional density of variables having indices from set K.

For a probability density function $f(x_K)$ and $L \subset K$ we can compute a marginal probability density $f(x_L)$ of $f(x_K)$ for almost every x_L in the following way

$$f(x_L) = \int_{\mathbf{X}_{K \setminus L}} f(x_K) dx_{K \setminus L}$$

where obviously the integration run over the domains of all variables in $K \setminus L$.

Having two sets of variable indices K and L, the symbol $f(x_{K \cap L})$ and $f^{\downarrow\{K \cap L\}}$ are two ways of corresponding marginal density representation. This density is marginalized up from $f(x_K)$, and thus represents a multivariate density of continuous random variables with indices from $K \cap L$.

For validity of definitions we require a notion of *support* (of a function), i.e., the set of points where function f has non-zero values

$$\operatorname{supp} f = \{x \mid f(x) \neq 0\}.$$

In case of densities, we mean positive values.

Having probability density $f(x_K)$ and two disjoint subsets $L, M \subseteq K$ we define *conditional probability density* of X_L given the occurrence of value $x_M \in X_M$ for almost every $x_{L \cup M}$ as

$$f(x_L \mid x_M) f(x_M) = f(x_{L \cup M}).$$

Let us remark that for $f(x_M) = 0$ the definition is ambiguous, but we do not need to exclude such cases.

3 Conditional independence

Continuous random variables x_K having a joint density $f(x_K)$ are all *independent* from each other if and only if for almost every x_K holds $f(x_K) = f_1(x_1) \cdots f_n(x_n)$. An important generalization of this notion is so called conditional independence.

Definition 1 (Conditional independence)**.** *Let us have probability density function $f(x_K)$ nad three disjoint subsets $L_1, L_2, L_3 \subseteq K$ where L_1 and L_2 are non-empty. We say that groups of variables X_{L_1} and X_{L_2} are* conditionally independent *given group X_{L_3} if almost everywhere*

$$f(x_{L_1 \cup L_2 \cup L_3}) f(x_{L_3}) = f(x_{L_1 \cup L_3}) f(x_{L_2 \cup L_3}). \tag{1}$$

We write $X_{L_1} \perp\!\!\!\perp X_{L_2} \mid X_{L_3}[f]$.

Let us mention the *principle of marginal zeroing*, i.e., if we have probability density $f(x_K)$ and $M \subseteq L \subseteq K$ then for almost every x_K holds $f(x_M) = 0$ implies $f(x_L) = 0$.

Now it is apparent that in Formula (1) the equality holds for x_K such that $f(x_{L_3}) = 0$. For almost every x_K such that $f(x_{L_3}) > 0$ we can according to definition above introduce conditional probability density and divide both sides of Formula (1) $f(x_{L_3})$ obtaining

$$f(x_{L_1 \cup L_2 \cup L_3}) = f(x_{L_1 \cup L_3}) f(x_{L_2} | x_{L_3}). \tag{2}$$

4 Operator of composition

The definition of the operator and its properties are in a way analogous to the situation in discrete probability distributions (see Jiroušek [3]).

Definition 2. *Consider two sets of continuous variables X_L and X_M, a probability density $f(x_L)$, and a probability density $g(x_M)$ with supports fulfilling the condition* $\operatorname{supp} f(x_{L \cap M}) \subseteq \operatorname{supp} g(x_{L \cap M})$. *The* right composition *is given by*

$$f(x_L) \triangleright g(x_M) = \frac{f(x_L) g(x_M)}{g(x_{L \cap M})}.$$

We can see that the composition exists when the assumption concerning the supports of the densities to be composed is fulfilled. The reason is that the definition in fact involves only the multiplication of density $f(x_L)$ by the conditional density $g(x_{M \setminus L} \mid x_{L \cap M})$, i.e., conditioned by the variables in the intersection. If there appears a zero in the denominator $g(x_{L \cap M})$ the condition on supports and principle of marginal zeroing imply that we get also product of two zeros in numerator and in this case we can quite naturally define the result of composition as 0.

Lemma 1. *For probability densities $f(x_L)$ a $g(x_M)$ such that*

$$\operatorname{supp} f(x_{L \cap M}) \subseteq \operatorname{supp} g(x_{L \cap M})$$

(thus the composition $f(x_L) \triangleright g(x_M)$ is defined) we can perform a marginalization with the result

$$(f \triangleright g)(x_L) = f(x_L)$$

almost everywhere.

Proof. First let us assume $g(x_{L\cap M}) > 0$. From the definition of marginalization we have

$$(f \triangleright g)(x_L) = \int_{\mathbf{X}_{M\setminus L}} (f \triangleright g)(x_{L\cup M}) dx_{M\setminus L} = \int_{\mathbf{X}_{M\setminus L}} \frac{f(x_L)g(x_M)}{g(x_{L\cap M})} dx_{M\setminus L}.$$

Now we can rewrite the density $g(x_M)$ using a conditional probability density in the following way.

$$(f \triangleright g)(x_L) = \int_{\mathbf{X}_{M\setminus L}} \frac{f(x_L)g(x_{L\cap M})g(x_{M\setminus L}|x_{L\cap M})}{g(x_{L\cap M})} dx_{M\setminus L} =$$
$$= f(x_L) \int_{\mathbf{X}_{M\setminus L}} g(x_{M\setminus L}|x_{L\cap M}) dx_{M\setminus L}$$

Since $g(x_{M\setminus L}|x_{L\cap M})$ is a conditional probability density we know that

$$\int_{\mathbf{X}_{M\setminus L}} g(x_{M\setminus L}|x_{L\cap M}) dx_{M\setminus L} = 1.$$

Now for $g(x_{L\cap M}) = 0$ we can use the assumptions on supports $\operatorname{supp} f(x_{L\cap M}) \subseteq \operatorname{supp} g(x_{L\cap M})$ which implies that also $f(x_{L\cap M}) = 0$. In this case we defined the result of composition as zero. □

Corollary 1. *From the preceding Lemma 1 for $X_M \subseteq X_L$ directly follows that*

$$f \triangleright g = f.$$

The result of composition is also a (multivariate) probability density in variables with indices from $L \cup M$, see the next Lemma.

Lemma 2. *For probability densities $f(x_L)$ a $g(x_M)$ such that* $\operatorname{supp} f(x_{L\cap M}) \subseteq \operatorname{supp} g(x_{L\cap M})$ *the composition $f(x_L) \triangleright g(x_M)$ is a probability density.*

Proof. Let us marginalize the result of composition $(f \triangleright g)(x_{L\cup M})$ over all variables

$$\int_{\mathbf{X}_{L\cup M}} (f \triangleright g)(x_{L\cup M}) dx_{L\cup M}.$$

This expression can be rewritten into two subsequent integrals where the inner one appeared in the proof of preceding lemma, i.e,

$$\int_{\mathbf{X}_{L\cup M}} (f \triangleright g)(x_{L\cup M}) dx_{L\cup M} = \int_{\mathbf{X}_L} \int_{\mathbf{X}_{M\setminus L}} (f \triangleright g)(x_{L\cup M}) dx_{M\setminus L} dx_L =$$
$$= \int_{\mathbf{X}_L} f(x_L) dx_L = 1.$$

The last integration follows from the fact that $f(x_L)$ is a probability density. □

Let us remark, that the definition of the operator of composition resembles the formulae of conditional independence (2). To be more precise, we can formulate the following assertion.

Lemma 3. *Let us have well defined composition* $h(x_{L\cup M}) = f(x_L) \triangleright g(x_M)$. *Then*

$$X_{L\setminus M} \perp\!\!\!\perp X_{M\setminus L} \,|\, X_{L\cap M}[h].$$

Proof. We have to show that for $h = f \triangleright g$ Formula (1) holds for almost every $x_{L\cup M}$.

For $x_{L\cup M}$ such that $g(x_{L\cap M}) = 0$ we have also $f(x_{L\cap M}) = 0$ thanks to the condition connecting supports of composed densities $\operatorname{supp} f(x_{L\cap M}) \subseteq \operatorname{supp} g(x_{L\cap M})$. This implies that also $h(x_{L\cap M}) = 0$ and from principle of marginal zeroing also $h(x_{L\cup M}) = h(x_L) = h(x_M) = 0$. Now we immediately see that for this case the assertion holds, because both sides of Equation 1 are equal to zero.

Now let us consider $x_{L\cup M}$ such that $g(x_{L\cap M}) > 0$. From Lemma 1 we know that marginal of composition $h(x_L) = f(x_L)$, now let us express $h(x_M)$ in a similar way as in proof Lemma 1

$$\begin{aligned} h(x_M) &= \int_{\mathbf{X}_{L\setminus M}} (f \triangleright g)(x_{L\cup M}) dx_{L\setminus M} = \\ &= \int_{\mathbf{X}_{L\setminus M}} \frac{f(x_L) g(x_M)}{g(x_{L\cap M})} dx_{L\setminus M} = \\ &= \int_{\mathbf{X}_{L\setminus M}} \frac{f(x_{L\cap M}) f(x_{L\setminus M}|x_{L\cap M}) g(x_M)}{g(x_{L\cap M})} dx_{L\setminus M} = \\ &= \frac{f(x_{L\cap M}) g(x_M)}{g(x_{L\cap M})} \int_{\mathbf{X}_{L\setminus M}} f(x_{L\setminus M}|x_{L\cap M}) dx_{L\setminus M} \end{aligned}$$

where

$$\int_{\mathbf{X}_{L\setminus M}} f(x_{L\setminus M}|x_{L\cap M}) dx_{L\setminus M} = 1.$$

Now we can express

$$\begin{aligned} h(x_L)h(x_M) = f(x_L)\frac{f(x_{L\cap M})g(x_M)}{g(x_{L\cap M})} &= \frac{f(x_L)g(x_M)}{g(x_{L\cap M})} f(x_{L\cap M}) = \\ = (f(x_L) \triangleright g(x_M)) f(x_{L\cap M}) &= h(x_{L\cup M}) h(x_{L\cap M}) \end{aligned} \tag{3}$$

where the last modification leading to term $h(x_{L\cup M})$ was done thanks to definition of composition and from Lemma 1 we know that $f(x_L) = h(x_L)$ which implies equality $f(x_{L\cap M}) = h(x_{L\cap M})$.

We arrived at equality $h(x_L)h(x_M) = h(x_{L\cup M})h(x_{L\cap M})$, i.e.,

$$X_{L\setminus M} \perp\!\!\!\perp X_{M\setminus L} \,|\, X_{L\cap M}[h].$$

□

Definition 3. *Two density functions $f(x_L)$ and $g(x_M)$ are* consistent *if $f(x_{L\cap M}) = g(x_{L\cap M})$.*

Note that densities $f(x_L)$ and $g(x_M)$ with $L \cap M = \emptyset$ are consistent. The pair of probability densities is commutative under assumption of consistency, see next Lemma.

Lemma 4. *For two consistent probability densities $f(x_L)$ and $g(x_M)$ such that both compositions $f(x_L) \triangleright g(x_M)$ and $g(x_M) \triangleright f(x_L)$ are defined holds*

$$f(x_L) \triangleright g(x_M) = g(x_M) \triangleright f(x_L).$$

Proof. From consistency assumption $f(x_{L\cap M}) = g(x_{L\cap M})$ and definition of operator of composition we directly see

$$f(x_L) \triangleright g(x_M) = \frac{f(x_L)g(x_M)}{g(x_{L\cap M})} = \frac{g(x_M)f(x_L)}{f(x_{L\cap M})} = g(x_M) \triangleright f(x_L).$$

□

Lemma 5. *Let us have two probability densities $f(x_L)$, $g(x_M)$ and P such that $L \cap M \subseteq P \subseteq L \cup M$*

$$(f \triangleright g)(x_P) = f(x_{L\cap P}) \triangleright g(x_{M\cap P}).$$

Proof. The composition $f \triangleright g$ is defined if and only if $f(x_{L\cap P}) \triangleright g(x_{M\cap P})$ is defined. Now let us marginalize the variables out in two consequent steps, first choose $M \setminus P$ and then $L \setminus P$ and let us rewrite both f and g using corresponding conditional distribution.

$$\begin{aligned}
(f \triangleright g)(x_P) &= \int_{\mathbf{X}_{L\setminus P}} \int_{\mathbf{X}_{M\setminus P}} \frac{f(x_L)g(x_M)}{g(x_{L\cap M})} dx_{M\setminus P} dx_{L\setminus P} = \\
&= \int_{\mathbf{X}_{L\setminus P}} \int_{\mathbf{X}_{M\setminus P}} \frac{f(x_L)g(x_{M\cap P})g(x_{M\setminus P}|x_{M\cap P})}{g(x_{L\cap M})} dx_{M\setminus P} dx_{L\setminus P} = \\
&= \int_{\mathbf{X}_{L\setminus P}} \frac{f(x_L)g(x_{M\cap P})}{g(x_{L\cap M})} \int_{\mathbf{X}_{M\setminus P}} g(x_{M\setminus P}|x_{M\cap P}) dx_{M\setminus P} dx_{L\setminus P} = \\
&= \int_{\mathbf{X}_{L\setminus P}} \frac{f(x_L)g(x_{M\cap P})}{g(x_{L\cap M})} dx_{L\setminus P} = \\
&= \int_{\mathbf{X}_{L\setminus P}} \frac{f(x_{L\cap P})f(x_{L\setminus P}|x_{L\cap P})g(x_{M\cap P})}{g(x_{L\cap M})} dx_{L\setminus P} = \\
&= \frac{f(x_{L\cap P})g(x_{M\cap P})}{g(x_{L\cap M})} \int_{\mathbf{X}_{L\setminus P}} f(x_{L\setminus P}|x_{L\cap P}) dx_{L\setminus P} = \\
&= \frac{f(x_{L\cap P})g(x_{M\cap P})}{g(x_{L\cap M})} = f(x_{L\cap P}) \triangleright g(x_{M\cap P})
\end{aligned}$$

The conditional densities marginalized out, since the integrals are equal to one. □

5 Compositional models

Analogously to the discrete case (see again Jiroušek [3]) we can iterate the operation of composition in order to build a multidimensional compositional model involving certain types of dependencies among variables. And the resulting multidimensional probability density is defined (if the assumptions of all operations hold) for all variables appearing at least once in the densities composed in the whole compositional model.

Let us stress that the operation of composition is generally neither commutative nor associative and the compositions are performed from left to right, i.e.,

$$f_1 \triangleright f_2 \triangleright f_3 \triangleright \cdots \triangleright f_{n-1} \triangleright f_n = (\ldots((f_1 \triangleright f_2) \triangleright f_3) \triangleright \cdots \triangleright f_{n-1}) \triangleright f_n.$$

Now let us present an assertion which under certain conditions allows to change the sequence of composition and comprises thus an associativity rule.

Lemma 6. *Let us have densities $f_1(X_{L_1})$, $f_2(X_{L_2})$ and $f_3(X_{L_3})$, now if $L_2 \subseteq (L_1 \cap L3)$ then*

$$(f_1 \triangleright f_2) \triangleright f_3 = f_1 \triangleright (f_2 \triangleright f_3).$$

The proof is completely analogous to the one presented in [3] for the discrete case.

6 Entropy of composition

Let us now explore the properties of the result of composition for probability densities from the viewpoint of information theory. First we will recall well-known definitions of differential entropy and Kullback-Leibler divergence (see, e.g., [4]).

Definition 4. *For multidimensional density $f(x_L)$ the* differential entropy $H(f)$ *is given by*

$$H(f) = -\int_{S_f} f(x_L) \log f(x_L) dx_L$$

where $S_f = \operatorname{supp} f(x_L)$.

Definition 5. *For two densities $f_1(x_L)$ and $f_2(x_L)$ the Kullback-Leibler divergence (or Kullback-Leibler distance, or relative entropy) $D(f_1 \parallel f_2)$ is defined by*

$$D(f_1 \parallel f_2) = \int f_1(x_L) \log \frac{f_1(x_L)}{f_2(x_L)} dx_L.$$

Notice that $D(f_1 \parallel f_2)$ is finite under the assumption that $\operatorname{supp} f_1 \subseteq \operatorname{supp} f_2$. For the sake of simplicity we can continuously set $0 \log \frac{0}{0} = 0$. Jensen's inequality (strict convexity of logarithm) implies that the Kullback-Leibler divergence is non-negative. The equality to zero occurs only if $f_1 = f_2$ almost everywhere.

The following pair of theorems was published in [1] together with both proofs.

Theorem 1. *Consider two sets of continuous variables X_L and X_M. If their corresponding densities $f(x_L)$ and $g(x_M)$ are consistent and their composition exists, the entropy of this composition is given by*

$$H(f \triangleright g) = H(f) + H(g) - H(g^{\downarrow L \cap M})$$

where $g^{\downarrow L \cap M}$ denotes marginal of density g with remaining variables $L \cap M$.

In Theorem 1 we expressed (under assumption of consistency) the entropy of composition using the entropies of particular operands. Now we will show that the operator of composition maximizes the entropy in the following sense.

Theorem 2. *For two sets of continuous variables X_L and X_M, and the corresponding consistent densities $f(x_L)$ and $g(x_M)$, consider a set of all common extensions of f and g denoted by $\Xi(f, g)$, i.e., a set*

$$\Xi(f, g) = \{h(x_{L \cup M}) | h(x_L) = f(x_L), h(x_M) = g(x_M)\}.$$

Then

$$f \triangleright g = \arg \max_{h \in \Xi(f,g)} H(h).$$

7 Dirac delta

In this section, we will introduce Dirac delta function as a degenerated density useful for practical realization of conditioning and intervention in compositional models. Another possible usage can be introduction of mixed models containing both discrete and continuous variables. In the preceding sentences we mentioned Dirac delta function, but in fact it is not a function. Dirac delta can be rigorously defined as a measure or as a distribution. In the theory of distributions a generalized function of Dirac delta is viewed not as a function but as the way how it affects other function (so called test function from some space) when integrated. Typically the space of test functions contains smooth functions (on real axis) with compact support (for more detail see, e.g., [6]).

For the sake of simplicity we will define Dirac delta function from the perspective of its properties. But first of all we recall Heaviside step function (or unit step function) in a variant

$$U(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0 \end{cases} \tag{4}$$

where symbol $U(x)$ was chosen in order not to confuse it with entropy.

Definition 6. *Dirac delta function is an object with the following properties*

Values: $\delta(x) = \begin{cases} +\infty & x = 0, \\ 0 & x \neq 0, \end{cases}$

Integration: $\int_{-\infty}^{\infty} \delta(x)dx = 1$,
moreover for any $\varepsilon > 0$ *it holds that* $\int_{-\varepsilon}^{\varepsilon} \delta(x)dx = 1$,

Sifting: *for any function* $f(x)$ *continuous in the* ε *neighborhood of* x_0 *so called sifting property holds*

$$\int_{-\infty}^{+\infty} f(x)\delta(x - x_0)dx = \int_{-\varepsilon}^{+\varepsilon} f(x)\delta(x - x_0)dx = f(x_0),$$

CDF: $\delta(x) = \frac{d}{dx}U(x)$, *where* $U(x)$ *is Heaviside step function (4).*

7.1 Conditioning

Now we will introduce the composition of Dirac delta function with continuous distribution as a mean of conditioning analogically to the degenerated probability mass function δ in discrete case in [2].

Theorem 3. *Let us have a probability density* $f(x_L)$, *single variable* $X \in X_L$ *and subset of variables* X_M *such that* $M \subset L$ *and* $X \notin X_M$. *Then*

$$f(x_M|x = x_0) = (\delta(x - x_0) \triangleright f(x_L))^{\downarrow X_M}$$

where again $\downarrow X_M$ *denotes corresponding marginalization.*

Proof. Denote $P \subset L$ such that $X_L = X_M \cup \{X\} \cup X_P$. Let us rewrite the righthand side of the equation in assertion

$$\begin{aligned}
(\delta(x - x_0) \triangleright f(x_L))^{\downarrow x_M} &= \int_{\mathbf{X}_{L\setminus M}} \delta(x - x_0) \triangleright f(x_L)dx_{L\setminus M} = \\
&= \int_{\mathbf{X}} \int_{\mathbf{X}_P} \delta(x - x_0) \triangleright f(x_L)dx_P dx = \\
&= \int_{\mathbf{X}} \int_{\mathbf{X}_P} \delta(x - x_0) f(x_{M\cup P}|x)dx_P dx = \\
&= \int_{\mathbf{X}} \delta(x - x_0) \int_{\mathbf{X}_P} f(x_{M\cup P}|x)dx_P dx = \\
&= \int_{\mathbf{X}} \delta(x - x_0) f(x_M|x)dx = f(x_M|x = x_0).
\end{aligned}$$

Where integration over variables X_P was simple marginalization and the last modification employed the sifting property. □

7.2 Intervention

Now let us take some preliminary consideration concerning modeling of intervention in case of continuous variables (for introduction and motivation concerning probabilistic approach to causality, interventions and do-calculus see [2] or [5]). According to

this we call compositional model with causal interpretation of dependencies among considered variables *causal compositional model*.

In agreement with the case of discrete compositional models [2] we a set of variables $X_K = \{X_1, \dots X_n\}$ where for any variable $X_i \in X_K$ set $\mathcal{C}(X_i) \subset X_K$ denotes its causes. We limit our considerations to the models in which variables can be ordered in such manner that causes precede their effects (for all k such that $X_k \in \mathcal{C}(X_i)$ must hold that $k < i$). Now denote $X_{L_i} = \mathcal{C}(X_i) \cup \{X_i\}$ and density $f_i(x_{L_i})$ describes the relation of variable X_i and its causes. So we consider causal model

$$h(x_K) = f_1(x_{L_1}) \triangleright f_2(x_{L_2}) \triangleright \cdots \triangleright f_n(x_{L_n}). \tag{5}$$

According to Pearl [5] the result of intervention on some variable $X \in X_K$ can be performed as a conditioning in a modified causal model where causal influence on the intervened variable is blocked, i.e., $\mathcal{C}(X) = \emptyset$. The intervened variable appears for the first time in ith density in the model and this new model is

$$h'(x_K) = f_i(X) \triangleright f_1(x_{L_1}) \triangleright f_2(x_{L_2}) \triangleright \cdots \triangleright f_n(x_{L_n}). \tag{6}$$

The expression is certainly valid for $i = 1$ thanks to the Corollary 1. For $i = 2$ it follows from the fact that neither $f_2(X)$ nor $f_1(x_{L_1})$ have any causes and we can apply Lemma 4 and then Corollary 1. In general case of $i > 2$ we also first apply Lemma 4, then $i - 2$ times Lemma 5 and finally Corollary 1 which results in

$$h'(x_K) = f_1(x_{L_1}) \triangleright f_2(x_{L_2}) \triangleright \cdots \triangleright f_i(X) \triangleright \cdots \triangleright f_n(x_{L_n})$$

where obviously $f_i(X)$ stands instead of $f_i(x_{L_i})$.

Theorem 4. *For the causal model h given by Formula 5, for any $X \in X_K$ and its value x_0 and for subset $X_M \subseteq X_K \setminus \{X\}$ we have*

$$h\big(x_M | do(x = x_0)\big) = \Big(\delta(x - x_0) \triangleright f_i(X) \triangleright f_1(x_{L_1}) \triangleright f_2(x_{L_2}) \triangleright \cdots \triangleright f_n(x_{L_n})\Big)^{\downarrow X_M}.$$

Proof. According to Pearl [5] the intervention is defined as a conditioning in the altered model (6), i.e., $h(x_M | do(x = x_0)) = h'(x_M | x = x_0)$. From this and Theorem 3 we get

$$h(x_M | do(x = x_0)) = \Big(\delta(x - x_0) \triangleright \big(f_i(X) \triangleright f_1(x_{L_1}) \triangleright f_2(x_{L_2}) \triangleright \cdots \triangleright f_n(x_{L_n})\big)\Big)^{\downarrow X_M}.$$

Now we use n times special associativity condition from Lemma 6 and finally Corollary 1 which leads to

$$h(x_M | do(x = x_0)) = \Big(\delta(x - x_0) \triangleright f_1(x_{L_1}) \triangleright f_2(x_{L_2}) \triangleright \cdots \triangleright f_n(x_{L_n})\Big)^{\downarrow X_M}.$$

□

This theorem shows the usage of Dirac delta in composition and presents the difference between conditioning and intervention.

8 Conclusion and further work

In the paper we presented an operator of composition for continuous random variables and analyzed its properties. Among others we elicited that the operator of composition embeds a relation of conditional independence and that it maximizes the entropy in certain sense. The results of the approach can serve as a computationally feasible model for approximate reasoning with continuous random variables allowing to interconnect low-dimensional densities without necessity to compute with the whole multidimensional density.

The Dirac delta was used to introduce conditioning and intervention to causal compositional models of continuous random variable in a consistent way. But it also provides a possibility to include deterministic variables, to define mixed random variables (partially continuous with steps in PDF) and to include discrete distributions to compositional models build from densities of continuous random variables which may be the aim of future work.

Acknowledgement
The research was supported by Grant Agency of the Czech Republic under project no. 15-00215S.

References

[1] Bína V. (2014), An Operator of Composition for the Multivariate Copula Construction. In *AIP Conference Proceedings* **1636**, MaxEnt'13, December 15-20, 2013, Canberra, Australia. pp. 5–10.

[2] Bína V. and Jiroušek R. (2015), On computations with causal compositional models. Accepted for publication in *Kybernetika*.

[3] Jiroušek R. (2011), Foundations of compositional model theory. *International Journal of General Systems* **40** (6), 623–678.

[4] Kullback S. and Leibler R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* **22** (1). 79–86.

[5] Pearl J. (2009). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, NY.

[6] Strichartz R. (1994). *A Guide to Distribution Theory and Fourier Transforms.* CRC Press, Boca Raton.

Reinforcement Structural Learning

Robert Brunetto
Department of Theoretical Computer Science and Mathematical Logic
Charles University in Prague, Faculty of Mathematics and Physics
robert@brunetto.cz

Marta Vomlelová
Department of Theoretical Computer Science and Mathematical Logic
Charles University in Prague, Faculty of Mathematics and Physics
marta@ktiml.mff.cuni.cz

Abstract

This article shows a novel approach to modelling and reinforcement learning of dynamic stochastic partially observable environment. We present an MCMC algorithm which learns the structure of a graphical model representing the environment. We use an approximation to a Bayesian method to learn posterior distribution over parameters of learned structure. The learning algorithm is online which allows us to use it in reinforcement learning setup. We demonstrate that this algorithm is usable on several simple experiments.

1 Introduction

Partially observable dynamic systems can be modelled using Hidden Markov model (HMM), which regards state as a single variable and observation as a single variable. Partially observable Markov decision processes(POMDPs) furthermore allow to place an agent in dynamic system and allow it to take actions and collect rewards. Modelling all possible states or all possible observations of a more complicated environment with one non-factorized variable could be impractical. That is why Dynamic Bayesian networks (DBNs) [1] are useful. They describe the environment by multiple variables.

Basic algorithms for inference in DBNs require that the structure of DBN is known and that values in conditional probability tables are exactly specified by an expert.

There are methods to learn DBNs from data [2]. Our algorithm differs from other existing DBN learning approaches in two points:

1. The structure and parameters are being learned dynamically as the agent collects observations.

2. Actions are selected in a way which tries to maximize collected reward.

According to the article [3] Ross and Pineau has successfully learned large structures of fully observable dynamic environments in 2012 by a MCMC method. Meanwhile Poupart and Vlassis presented a novel way to learn the parameters of DBN representing partially observable environment.

Both of these articles utilize a Bayesian approach and we argued in a recent article [4] that they are suitable for combining to a single algorithm which learns both structure and parameters in a partially observable environment.

In this article we repeat the main ideas, explain them in more details and report first experimental results.

2 Proposed model

We suppose that environment behaves as DBN with the exception that one of the variables denoted A is not random but selected by the agent. We do not assume that the structure of this graphical model is known. The environment is in an partially observed state $\mathbf{s} \in S$ at each time step, where S denotes the state space. We assume that it can be factorized according to state variables $X \in \mathbf{X}$ hence each state $\mathbf{s}$ is a vector $\mathbf{s} = (s_1, ...s_{|\mathbf{X}|}) \in S = \prod_{X \in \mathbf{X}}$.

We assume that some of the variables from $\mathbf{X}$ may be observable. We denote them $\mathbf{O} \subset \mathbf{X}$. We refer to the part of the state which is observed as observation and we denote it $\mathbf{o}$. It is actually $\mathbf{s}$ restricted to $\mathbf{O}$ which we denote as $\mathbf{o} = \mathbf{s}_\mathbf{O}$. Other restriction operation will be denoted analogously.

We always restrict vectors which are denoted by bold lower case letters to sets of variables, which are denoted as bold upper case letters. Capital letter which are not bold will denote single variable. Lower case non-bold letter denotes its value.

Specially in case when G is the graph of a Bayesian network, $\mathbf{PA}_G X$ denotes the set of parents[1] of variable X in graph G. Values of variables of parents of X will be denoted $\mathbf{pa}_G X$ where $X \in \mathbf{X}$.

When the time advances to the next step the state $\mathbf{s}$ changes to $\mathbf{s}'$ according to unknown transition probability $P(\mathbf{s}'|\mathbf{s})$.

We assume that the transition probability $P(\mathbf{s}'|\mathbf{s}, a)$ factorizes according to a dynamic Bayesian network with unknown structure G with unknown parameters.

$$P(\mathbf{s}'|\mathbf{s}, a) = \prod_{X \in \mathbf{X}} P(\mathbf{s}'_X|(\mathbf{s}', \mathbf{s}, a)_{\mathbf{PA}_G X}) \tag{1}$$

Dynamic Bayesian network with structure G is described as Bayesian network which has $\mathbf{X}' \cup \mathbf{X} \cup \{A\}$ as its nodes. By $(\mathbf{s}', \mathbf{s}, a)_{\mathbf{PA}_G X}$ we denote $(\mathbf{s}', \mathbf{s}, a)$ restricted to the variables which are parents of variable X according to structure G.

To emphasize that we take $P(\mathbf{s}'_X|(\mathbf{s}', \mathbf{s}, a)_{\mathbf{PA}_G X})$ as an unknown parameter, we denote it:

[1] Parent variable V of variable W is a term used in Bayesian-networks-literature to denote that there is an arrow from V to W in the graph of Bayesian network.

$$\theta_X^{(\mathbf{s}',\mathbf{s},a)_{\mathbf{PA}_G X}}. \tag{2}$$

It is actually a vector of real values between 0 and 1 summing to 1 containing for each value $\mathbf{s}'_X$ the probability $P(\mathbf{s}'_X|(\mathbf{s}',\mathbf{s},a)_{\mathbf{PA}_G X})$. As the indexes suggest that we have such a set of parameters for each state variable X and for each combination of values of its parents.

Even though we do not assume structure G to be known we still assume that a prior probability distribution $P(G)$ over structures G is known. This probability distribution can express expert's prior knowledge or it can just prefer simple structures over more complicated ones.

The transition probability can be expressed as:

$$P(\mathbf{s}'|\mathbf{s},a) = \sum_G P(G) \cdot P(\mathbf{s}'|\mathbf{s},a,G). \tag{3}$$

We assume that the parameters $\theta_X^{(\mathbf{s}',\mathbf{s},a)_{\mathbf{PA}_G X}}$ follow the Dirichlet distribution, which is conjugate to multinomial distribution.

Each unknown parameter $\theta_X^{(\mathbf{s}',\mathbf{s},a)_{\mathbf{PA}_G X}}$ can be regarded as an additional state feature. This way a DBN with unknown parameters can be converted into a bigger Bayesian network without unknown parameters. From this point of view, learning the state is equivalent to learning the state and the dynamics of the environment. During the learning process we maintain probability distribution over possible states. We call this *belief*.

Surprisingly, as Poupart and Vlassis showed [5], even though there is an infinite number of possible parameters the belief for a given structure can be maintained in a closed form. We will review it in section 2.1.

There is only a finite number of possible graphical structures implying that the belief over all of them can be maintained in a closed form. But the number of possible graphs may be large. That is why approximation is introduced in section 2.2.

2.1 Belief for a given structure

We begin by describing how belief looks like when we are given the structure. It is exactly the same as described by Poupart & Vlassis [5].

The key component is the nice properties of Dirichlet distribution. Let us have one discrete random variable V which can take values from 1 up to K with probabilities $(\theta_1, \theta_2, ...\theta_K)$, where $\sum_i \theta_i = 1$.

The usual way to estimate these parameters in Bayesian statistics is to compute the posterior when assuming that the prior follows Dirichlet distribution. Its density is given by $D(\theta;\mathbf{n}) = \frac{1}{B(\mathbf{n})} \prod_i \theta_i^{n_i-1}$, where $\theta = \{\theta_i\}_i$, $\mathbf{n} = \{n_i\}_i$ are some hyperparameters and $B(\mathbf{n})$ is a constant depending on them which makes the distribution sum to one. It is known as multinomial beta function.

The prior distribution which states that we have no evidence is expressed by setting all hyperparameters n_i equal to 1. It can be interpreted as evidence that all values were observed exactly once or it can be thought of only as smoothing the posterior.

The posterior probability that the variable V contains value i is then equal to proportion of how many times was value i observed.

$$P(V = i) \quad = \quad \frac{n_i}{\sum_i n_i}. \tag{4}$$

In the fully observable environment then we could estimate all parameters in the whole structure by (4). We would use this estimate for each state variable X and for each combination of values of its parents $\mathbf{pa}_G X$. That is the reason why we add X as a lower index to θ and $\mathbf{pa}_G X$ as an upper index to θ as in (2).

Each set of parameters $\theta_X^{\mathbf{pa}_G X}$ sums to one.

From now on θ without any indexes will denote all these sets of parameters together.

In the simplified case when the whole history is observed the density of probability of being in "information state" θ is

$$\prod_{X,\mathbf{pa}_G X} D(\theta_X^{\mathbf{pa}_G X}; \mathbf{n}_X^{\mathbf{pa}_G X}). \tag{5}$$

The problem that not all state features are observable can be overcome by the following theorem proven by Poupart and Vlassis [5].

Theorem 1. *If the prior is a mixture of products of Dirichlets*

$$b(\mathbf{s}, \theta) = \sum_i c_{i,\mathbf{s}} \prod_{X',\mathbf{pa}_G X'} D(\theta_{X'}^{\mathbf{pa}_G X'}; \mathbf{n}_{X',i,\mathbf{s}}^{\mathbf{pa}_G X'}) \tag{6}$$

then the posterior is also a mixture of products of Dirichlets

$$b_{o'}(\mathbf{s}', \theta) = \sum_j c_{j,\mathbf{s}'} \prod_{X',\mathbf{pa}_G X'} D(\theta_{X'}^{\mathbf{pa}_G X'}; \mathbf{n}_{X',j,\mathbf{s}'}^{\mathbf{pa}_G X'}). \tag{7}$$

But how can we get the probability of being in a specific state from this representation? We show in the following theorem that this can be easily done.

Theorem 2. *Parameter θ in formula (6) for the mixture of products of Dirichlet can be integrated out in a closed form.*

Proof.

$$b(\mathbf{s}) \quad = \quad \int b(\mathbf{s}, \theta) = d\theta \tag{8}$$

$$\quad = \quad \sum_i c_{i,\mathbf{s}} \int \prod_{X',\mathbf{pa}_G X'} D(\theta_{X'}^{\mathbf{pa}_G X'}; \mathbf{n}_{X',i,\mathbf{s}}^{\mathbf{pa}_G X'}) d\theta = \tag{9}$$

$$= \sum_i c_{i,\mathbf{s}} \prod_{X',\mathbf{pa}_G X'} \int D(\theta_{X'}^{\mathbf{pa}_G X'}; \mathbf{n}_{X',i,\mathbf{s}}^{\mathbf{pa}_G X'}) d\theta = \tag{10}$$

$$= \sum_i c_{i,\mathbf{s}} \tag{11}$$

The equation (8) defines symbol $b(\mathbf{s})$. The equation (9) follows from the definition of $b(\mathbf{s}, \theta)$ by switching sum and integral. The equation (10) switches integral and multiplication which is possible in this case because each factor depends on the different variable of multidimensional integration. Density of all probability distributions (including Dirichlet distribution) integrates to one and the product of ones is one. That is why the equation (11) holds. □

Despite this encouraging results the number of components in the mixture (7) grows exponentially. Luckily the belief can be approximated by the approximation proposed by Poupart & Vlassis. This will be described in section 2.4. Firstly, in the next section, we describe the way the belief over possible structures is maintained.

2.2 Belief over structures

The simplest and most naive approach to maintaining the overall belief which contains the probability of structure, its parameters and state is straightforward. It is sufficient to keep the belief for each structure and the probability of the structure. The problem is that the number of graphs on given number of vertices grows very fast with the increasing number of vertices but we want to maintain only small number of graphs.

We propose to remember only one randomly chosen structure where the probability that the structure G is chosen would be proportional to the probability $P(G|history)$.

The probability $P(G|history)$ could be difficult to compute. But, as Ross & Pineau noted in article [3] about MDP, a Markov chain of graphs can be maintained using Metropolis-Hastings algorithm. The algorithm will ensure that the Markov chain converges to a distribution of graphs which is equal to $P(G|history)$.

The Metropolis-Hastings algorithm needs to use $P(history|G)$ which can be computed as follows: It is equal to $P(\mathbf{o}|b, G) \cdot P(h_{t-1}|G)$, where h_{t-1} denotes history up to the previous time step.

Then the idea of computing $P(\mathbf{o}'|b, a)$ is as follows: $P(\mathbf{o}'|b, a)$ is equal to sum of $P(\mathbf{s}'|b,)$ over states $\mathbf{s}'$ compatible with observation $\mathbf{o}'$. $P(\mathbf{s}'|b, a)$ can be computed directly from hyperparameters. States s' compatible with observation o are enumerated during belief update procedure which converts one mixture of products of Dirichlets (6) to other mixture of products of Dirichlets (7). Assuming that total weight of components of mixture for belief in time $t-1$ is $P(history_{t-1}|G)$, the total weight of samples in time t is $P(history_t|G)$. Hence the algorithm for computing $P(history_t|G)$ is straightforward.

We propose to start with any arbitrary structure, then learn its parameters by the application of theorem 1 and approximations from next section. Then we propose to switch the structure after specified amount of time and repeat the whole process.

The algorithm for switching the structure is given in algorithm 1 where $q(G'|G)$ is the probability of transition from graph G to graph G'. This distribution can be set any arbitrary way which will ensure that all graphs are reachable. $P(G)$ is prior distribution over graph structures and the probability $P(history|G')$ can be computed as described above.

Random transitions with probability $min\left(1, \frac{P(history|G')P(G')q(G'|G)}{P(history|G)P(G)q(G'|G)}\right)$ are well known under the name Metropolis-Hastings algorithm. This algorithm ensures that the Markov chain of graphs converges to the distribution $P(G|history)$ which implies that our algorithm eventually learns either the correct structure or a structure which is equally good with respect to encountered history.

We implemented random changes and distribution $q(G'|G)$ as local change of the graph structure by deleting or adding an edge.

This change depends on distribution $q(G'|G)$ and isn't explicitly written in algorithm 1. We assume that this change is done on the line $G := G'$ which changes the structure.

Algorithm 1: switchStructure(G)

Data: structure G
Result: possibly modified structure G
G' := random modification of G;
With a probability $min\left(1, \frac{P(history|G')P(G')q(G'|G)}{P(history|G)P(G)q(G'|G)}\right)$
    G:=G';

2.3 Action selection

Agent's goal is to collect as much reward as possible. It needs to select best combination of actions to do so. Mappings from believes to actions are called policies. The question how to find best or good policies for known models is intensively researched and information can be found in POMDP-related literature.

Our algorithm learns the distribution of possible models. As an approximation, the best model can be selected and then any of the existing POMDP solvers can be used to find the policy and select action.

There even exist an approach which can use distribution of possible models in the form of product of Dirichlets to construct a policy [5]. We suggest to use such techniques off-line after the model has been learned. During learning fast action selection algorithm needs to be used. We suggest to select actions randomly at the beginning when there isn't known information about parameters of structure and hence when there is none or only small amount of information about consequences of actions. Afterwards we suggest to select action which appears the best when compared in simple simulation using learned models.

2.4 Approximate representation of the mixture of products of Dirichlets

As noted in section 2.1 the number of components of the mixture representing current belief for a given graph grows exponentially with time which is untraceable. Each component of the mixture is associated with coefficients $c_{i,\mathbf{s}}$. Naturally some of these coefficients will be smaller while the others will be bigger. We propose to handle this approximately and sample only some components with bigger coefficients.

For description of possible approximations along with some of their advantages and disadvantages we encourage the reader to read [5] where the same approximations are described.

3 Experiments

We tested our approach on several synthetic experiments. In each experiment we approximated mixture of product of Dirichlets by 100 samples. The algorithm switched the structure each 800 time steps and we let the algorithm switch 1500 structures. Then we learned again the best structure for 2000 time steps to compare the results.

We include precise description of all experiments including the description of transition models.

3.1 Water levels modelling

Our first experiment is based on practical motivation. It models behaviour of river's water levels. Possible usage of such model could be for example floods prediction. However our goal is not modelling reality but instead watching what the model learned and allowing to learn everything from scratch. That is why we don't use expert information already known about this problem. We also want to compare the learned model with the model generating data. That justifies why we use synthetic data rather than real data.

We simplified reality as follows: Two variables are observed, $Rain$ which indicates whether it is raining or not and variable $WaterLevel$ indicating water levels. For simplicity we assume only one hidden variable $ReservoirFullness$ indicating the state of hidden natural water reservoirs.

In our data generating model the rain behaves as follows: If it has rained, the probability of the rain in the next day is 0.8 and if it did not rain during the previous day the probability of raining during the next day is 0.2. We also assume that the fullness of water reservoir behaves randomly and that it depends only on the previous value of itself. The probability that the fullness of water reservoir will not change is 0.9.

The water level stochastically depends on the $Rain$ variable and on the $ReservoirFullness$ variable as follows: If the reservoir is full, the rain causes higher water levels more likely, not raining causes lower water levels more likely. In the opposite case, when the water reservoir is not full, we let the rain have exactly opposite influence on water level. All relevant probabilities were set to 0.9 or 0.1 respectively.

3.2 Telephone

The previous experiment is over a fixed number of variables and cannot be scaled over different number of variables. That is why we introduce the next experiment. Telephone (or Chinese whispers in British English) is a game played around the world, in which one person whispers a message to another, the message passed through a line of people until the last player announces it to the entire group. In our experiment we represent each player by a variable. First player generates random messages consisting of one 0 or one 1 with probability 50% at each time step. Other variables (players) copy message from previous player at each time step. Last player in a row is our agent. His task is to tell what the value of previous variable was at previous time step. The agent can tell the answer through action and it gets reward 1 whenever it answers correctly and reward 0 if it answers incorrectly.

To test how our algorithm scales with increasing number of variables we tried this experiment multiple times with various number of players. To test that agent is capable of using hidden variables and then learn the structure we made every second variable unobserved. Which means that agent does not receive information from previous player but instead it observes value of second last player. The correct strategy for the agent is remembering what the message is (in a hidden variable) and then answering it two time steps later.

In our first setup we assumed that the model behaves deterministically - players forward messages without errors. If i-th variable is observable then its value at time $t + i - 1$ equals the value of first variable at time t.

We tested the telephone structure with 3,5,7, 9 and 11 variables. We also tested non-deterministic variant of this experiment. If i-th variable is observed at time step t then the same value is observed in $i+2$-th variable with probability 0.85 and opposite value with probability 0.15 at time step $t + 2$.

We expected that successfully learned structures should contain arrows from i-th variable in one time slice to $i+1$ in the next time slice, but possibly with permutated hidden variables and with values 0 and 1 flipped and with unnecessary arrows. To our surprise the program found other models correctly representing the situation. They were composed of several variables with XOR-like behavior with arrows inside one time slice which were used to transfer information in the opposite direction of the arrow.

3.3 Practical note

Metropolis-Hastings algorithm can be used to create Markov chain of samples from any distribution which one knows up to scaling constant. This is used in our work to sample Markov chain of structures which are sampled according to $P(G|history)$. Despite nice theoretical properties, mixing time is also very important property in practice.

Consider that Metropolis-Hastings algorithm compares structures G_1 and G_2. The typical probability of observation given structure G_1 is 0.55. The structure G_2 performs slightly worse and typical probability of observation is 0.50. Then assuming

that history consist of 800 observations, the ratio
$P(history|G_1)/P(history|G_2)$ would be equal to $0.55^{800}/0.50^{800} = 10^{33}$. The correct conclusion is that structure G_2 is 10^{17} times worse that G_1. Although structure G_1 might be much better than G_2, it could still be only local optimum and it could be necessary that the Markov chain of structures go to G_2 and then by modification of G_2 it will go to some structure which performs even better than G_1. Hence even though the stationary distribution of this Markov chain follows $P(G|history)$, it could take 10^{17} steps until the chain moves from local optimum to a structure with 5% smaller observation probabilities.

To decrease the mixing time we replaced $P(G|history)$ by $P(G|history)^{\frac{12}{800}}$ which intuitively means that we take probabilities of 800 observations as if that were probabilities only of 12 observations. This decreased mixing time to reasonable values while preserving the property that the Markov chain stays in good structures for most of the time.

3.4 Comparing the results

We compared the theoretically computed entropy of model generating data against the approximate log-likelihood of best structure reported by the algorithm. The reported log-likelihood of selected structure can be overfitted and choosing the best structure could cause that the most overfitted structure is selected. To see how big this influence is, we learned again the parameters on the best structure now for 2000 time step. The result is in column *val. data* in table 1. Column *best* shows what log-likelihood was reported during learning the parameters for given structure on 800 observations. Column *limit* shows what is the log-likelihood of data given the generating model.

Column *time* shows how many structures the program tried before model exceeded specified log-likelihood. The specified log-likelihood depends on complexity of the environment. It is equal to $(2 * limit + trivialLimit)/3$ where $trivialLimit$ is log-likelihood of structure without arrows.

In our experiments, the agent at first doesn't know what actions to take to get reward. So we expected that agent will randomly get reward 1 with probability 0.5 and reward 0 with probability 0.5. Hence the average amount of reward received will be for some time around 0.5. The chance to get reward increases after the search through possible structures finds necessary arrows first time. The experiments confirmed this conjecture. To exclude this initial randomness from table we measured the time when the agent started collecting rewards. This is shown in column *rw start* and then we averaged the rewards from that point onwards in column *rw*.

We started all experiments 20 times. Results in table 1 are averages from all the runs. However However in some cases the program did not find the structure which exceeded specified log likelihood in given time limit (which is 1500 structure changes). Hence in some cases we did not have values to average in the column *time*. In the column *f* (as fail) we report the number of such cases. The column *time* reports the average of all other cases.

The column *r rw* shows the maximal fraction of reward collected by best structure.

The column *r ll* shows the fraction of reward collected by structure with best log-likelihood. The maximal possible value in the last two columns is 1. However in non-deterministic case the expected value in the non-deterministic case for the best possible action selection policy is 0.85.[2]

#vars	limit	best	val. data	time	rw start	rw	f	r rw	r ll
floods									
3	-1.5	-1.3	-1.53	1.12	-	-	0	-	-
telephone									
3	-1.00	-1.0	-1.01	78.05	49.9	0.88	0	1.00	0.99
5	-1.00	-1.0	-1.02	235.0	87.0	0.87	0	1.00	0.99
7	-1.00	-1.3	-1.70	586.2	321.1	0.67	1	0.94	0.88
9	-1.00	-1.8	-1.87	450.8	536.4	0.71	0	0.88	0.74
11	-1.00	-2.3	-2.55	646.4	872.1	0.84	1	0.78	0.73
nondeterministic telephone									
3	-1.61	-1.4	-1.62	31.79	25.6	0.71	0	0.87	0.86
5	-2.22	-2.1	-2.32	451.4	179.9	0.64	0	0.87	0.82
7	-2.83	-2.9	-3.19	392.1	311.2	0.58	0	0.85	0.74
9	-3.44	-3.8	-4.13	598.1	501.6	0.57	2	0.81	0.66
11	-4.05	-4.2	-4.56	603.7	1006.6	0.52	1	0.65	0.52

Table 1: Results of experiments

3.5 Interpretation of results

You can see that the best model reported by the algorithm is slightly overfited when compared on validation data (columns *best* vs. *val. data*). But even though learned models perform well on validation data. You can compare column *val. data* with column *limit* to see that learned models are similar to models generating data.

Sadly, time needed to learn the model grows with number of variables. As can be seen from the table, the time before good model appeared and time when the agent started collecting rewards grows with number of variables (columns *time* and *rw start*).

The amount of reward the agent would collect when using learned model (column *r rw*) is surprisingly high. But the amount of reward that agent receives during learning *r ll* is better than random but still far from perfect. This is caused by random changes of structure which are not always good but sometimes bad and hence cause geting smaller reward.

[2] Values 0.86 and 0.87 imply that some overfitting did happen. Agent most certainly found right structure and policy, then maximization over them selected the case in which the agent was most lucky.

3.6 Best likelihood vs. best reward

The experiments confirm that algorithm is usable in some environments. Found models should converge to models with high likelihood.

Even our simple experiments show significant difference between best reward collecting structures and highest likelihood structures (columns *r rw* and (t ll)). This can be easily explained: The complete model is not required to be known in order to select right action in these experiments.

It seams that making the model better cannot make any harm. But suppose the following scenario: The agent is maximizing the likelihood of model. It can rise to local or global optimum or it can get stack in a space of models with similar or equal likelihoods. We even believe that there could exist models which are equally likely or almost equally likely as data generating model. (Imagine the telephone experiment in which each player flips the message for an instance.)

If the agent learns model with switched values of hidden variables it won't make any change. But in other experiments it could happen that two or more models are equally likely but using one leads to greater reward than using the other.

Unfortunately approximations will select only few of high-likelihood models and can miss other high-likelihood model. When the agent chooses action according to model with high likelihood it does not necessarily imply that it will also get high reward.

This simple debate shows that learning reward yielding models rather than high likelihood models is also important. We consider this as our future work.

3.7 Conclusion and future work

Mixtures of products of Dirichlets can be successfully utilized to learn the parameters of DBN describing the partially observable environment. The Metropolis-Hastings algorithm can be used to switch between structures and search for the best ones. The experiments we performed show encouraging results and that this approach is applicable and that it can be used in reinforcement learning.

The overall performance depends on several parameters set by user and we do not give an answer to how these parameters should be tuned.

Namely the following ones (which were set during experiments to following values): The number of samples in the mixtures of products of Dirichlets (100), the time for learning parameters of each structure (800), the constant from section 3.3 balancing mixture time with quality of learned structures (12), the prior distribution over structures ($P(G) \propto 0.85^{\#parameters}$), proposal function q from Metropolis-Hastings algorithm (uniform probability between structures differing by one edge).

We expect that after tuning the mentioned parameters, even better results could be reached.[3] However, adapting these parameters automatically during the learning process would be much more interesting. We consider this as our future work.

[3] We consider our implementation just a proof of concept. We assume that by a more careful implementation of approximations from [3] the program could be speeded up significantly.

Our approach to learning structures of partially observable dynamic environments is also innovative by the fact that it is online. It doesn't need all observations to be known before the program is started and actual model can be used any time to select actions.

The experiments shows not only usability and scalability of proposed approach but we also discuss interesting phenomenon that highest likelihood model needs not to be the best ones for action selection. This might be related but is not equivalent to well known exploration-exploitation dilemma.

Acknowledgments.

This research was supported by SVV project number 260 104 and by project number P103-15-19877S.

References

[1] Murphy, K. P. (2002) *Dynamic bayesian networks: representation, inference and learning*. Ph.D. thesis, University of California, Berkeley.

[2] Friedman, N., Murphy, K., and Russell, S. (1998) Learning the structure of dynamic probabilistic networks. *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 139–147, Morgan Kaufmann Publishers Inc.

[3] Ross, S. and Pineau, J. (2012) Model-based bayesian reinforcement learning in large structured domains. *arXiv preprint arXiv:1206.3281*.

[4] Brunetto, R. and Vomlelová, M. (2013) Acting and bayesian reinforcement structure learning of partially observable environment. *Informacne Technologie-Aplikacie a Teoria*, **13**, 8.

[5] Poupart, P. and Vlassis, N. A. (2008) Model-based bayesian reinforcement learning in partially observable domains. *ISAIM*, Citeseer.

A learning methodology for coherent hybrid probabilistic fuzzy classifiers

Andrea Capotorti
Dept. Matematica e Informatica
University of Perugia
andrea.capotorti@unipg.it

Davide Petturiti
Dept. Matematica e Informatica
University of Perugia
davide.petturiti@unipg.it

Valentina Poggioni
Dept. Matematica e Informatica
University of Perugia
valentina.poggioni@unipg.it

Abstract

We aim at redesigning the hybrid fuzzy classifier proposed in [8] that joins together probabilistic inference with classical Wang-Mendel fuzzy rule bases. We will profit from coherent probabilistic fuzzy IF-THEN rules, as already described in [3], with a novel elicitation strategy based on a new learning methodology. This will lead us to propose a probabilistic fuzzy rule based classification algorithm. The methodology for constructing and drawing inferences from a probabilistic fuzzy rule based classifier guarantees the global coherence of the probability evaluations and allows to take into account potentially imprecise (lower-upper) probabilistic conclusions. The proposed classification algorithm will be tested on a doping alert problem and compared with two other fuzzy IF-THEN rule based classifiers on artificial datasets.

Keywords: Probabilistic fuzzy system, probabilistic fuzzy IF-THEN rule based classifier, coherent conditional probability

1 Introduction

This paper is a first proposal for constructing hybrid probabilistic fuzzy classifiers based on the coherent conditional probability paradigm. Due to space limitations, our goal here is to sketch the main steps to reach the goal, leaving implementation details to future contributions.

Conciliation between probability and fuzzy theories has been largely debated in literature and several approaches have been proposed, leading to a plethora of different hybrid methodologies (see, e.g., [10, 11, 12, 13, 14, 16, 17]).

Even though the question of consistence with a framework of reference is of prominent importance for an hybrid methodology to produce meaningful results, the existing proposals do not seem to pay particular attention to this fact.

In the recent past the two theories have been consistently joined together thanks to the Coletti & Scozzafava representation of fuzzy sets through coherent conditional probabilities (see, e.g., [5, 6, 7]). Hence, it is possible to complete the aforementioned hybrid models by integrating them through probabilistic techniques, all maintaining consistence with the coherent conditional probability framework.

In particular, we focus on classification problems where both uncertainty and vagueness are present. A prototypical real problem of this kind is the doping diagnosis, or better "alert", where a claim on the suspect use of forbidden drugs by a non-professional athlete must be performed. Indeed, the doping alert problem is characterized by the presence of linguistic descriptions of some attributes (such as increasing muscles, severe headaches, high blood pressure, and so on) together with the possible absence of a specific test (the so-called "gold standard") on some quantitative feature.

The paper sketches a procedure to build a probabilistic fuzzy IF-THEN rule based classifier starting from a training set of examples, each endowed with a linguistic class label. Differently from other proposals present in the literature (see, e.g., [8]), our method uses the training set also to estimate the membership functions of fuzzy antecedents and consequents of IF-THEN rules. In particular, all the probabilistic evaluations that are produced during the construction phase are coherent conditional probabilities. Once the rule based classifier is built, the paper discusses its use in a classification task starting from an instance whose description is either quantitative (crisp) or linguistic (fuzzy). The proposed classification algorithm is first applied to a doping alert problem showing its good robustness with respect to the choice of the t-norm used to realize fuzzy operations. Then it is compared with the Wang-Mendel classifier and with the classifier described in [8] in a classification task on artificial datasets.

The paper is organized as follows. Section 2 introduces the main tools and the theoretic interpretation of probabilistic fuzzy IF-THEN rules. Section 3 specifies the practical construction of the rule base. Section 4 describes the reasoning mechanism, dividing the situations where the new case to classify is expressed by crisp or fuzzy attribute values. Section 5 illustrates an application to a doping alert problem and some preliminary results on an empirical study, finally, Section 6 concludes by giving some outline on future developments.

2 Preliminaries

In line with Coletti & Scozzafava interpretation [5, 6, 7], we intend a generic fuzzy set connected to a linguistic characteristic φ of a random variable (r.v.) X with range χ, as a couple of the form

$$A^{\star}_{\varphi} = (E_{A^{\star}_{\varphi}}, \mu_{A^{\star}_{\varphi}}(x)) \tag{1}$$

where $E_{A^\star_\varphi}$ denotes the Boolean event "You claim that X has property $A^\star_\varphi$", while $\mu_{A^\star_\varphi}(x) : \chi \to [0,1]$ is the associated membership function intended as the coherent conditional probability $\mu_{A^\star_\varphi}(x) = P(E_{A^\star_\varphi}|X = x)$ and expressing the (subjective) probability of classifying with linguistic label $A^\star_\varphi$ an instance showing the crisp value x for the r.v. X. Let us recall that every assessment $P(E_{A^\star_\varphi}|X = x)$, for $x \in \chi$, ranging in $[0,1]$ and such that $P(E_{A^\star_\varphi}|X = x) = 0$ if $E_{A^\star_\varphi} \wedge (X = x) = \emptyset$ and $P(E_{A^\star_\varphi}|X = x) = 1$ if $(X = x) \subseteq E_{A^\star_\varphi}$, is a coherent conditional probability [5]. In the sequel we denote with $\mathfrak{F}_\chi$ the set of all fuzzy sets on χ according to (1) and we avoid to explicitly write the linguistic characteristic φ of a fuzzy set in order to have a lighter notation.

Instances of the classification problem are described by *a attributes*, i.e., by r.v. $X_1, \ldots, X_a$ which are supposed (at present stage) to be logically independent and so the random vector $(X_1, \ldots, X_a)$ has the Cartesian product $\chi_1 \times \ldots \times \chi_a$ as range. For each attribute X_i, $i = 1, \ldots, a$, classification *instances* - also named *profiles* - can show either crisp $x_i \in \chi_i$ or fuzzy $A^\star_i \equiv (E_{A^\star_i}, \mu_{A^\star_i}(x_i)) \in \mathfrak{F}_{\chi_i}$ values.

In the following we will use bold letters to denote multidimensional quantities, hence the random vector of attributes will be denoted as $\mathbf{X} = (X_1, \ldots, X_a)$, $\mathbf{x} = (x_1, \ldots, x_a)$ will indicate a crisp point in its range $\chi_1 \times \ldots \times \chi_a$, while $\mathbf{B}^\star$ a fuzzy set in $\mathfrak{F}_{\chi_1 \times \ldots \times \chi_a}$ with membership function $\mu_{\mathbf{B}^\star}(\mathbf{x})$.

Due to the vagueness in the classification task, we consider class labels $C^\star_j$, $j = 1, \ldots, d$, as fuzzy linguistic labels, hence we operate as in Mamdani-type fuzzy systems. In general, classes $C^\star_j$ refer to an output variable Y with range Υ, where it can be $Y = \mathbf{X}$ and so $\Upsilon = \chi_1 \times \ldots \times \chi_a$. Since, in general, it is not easy to determine an output variable Y that can be used as target of the classification and that is distinct from $\mathbf{X}$, in the present paper we assume $Y \equiv \mathbf{X}$, thus linguistic class labels are denoted as $\mathbf{C}^\star_j$, $j = 1, \ldots, d$, since they belong to $\mathfrak{F}_{\chi_1 \times \ldots \times \chi_a}$.

Probabilistic fuzzy classifiers are composed by a set of rules that have the following general structure [12]:

$$\text{IF } \mathbf{A}^\star_i \text{ THEN } \mathbf{C}^\star_j \text{ with probability } w_{j|i}, \quad j = 1, \ldots, d, \tag{2}$$

where $\mathbf{A}^\star_i$ is an "antecedent" fuzzy set for the random vector of attributes $\mathbf{X} = (X_1, \ldots, X_a)$, i.e., $\mathbf{A}^\star_i \in \mathfrak{F}_{\chi_1 \times \ldots \times \chi_a}$, and $w_{j|i}$ are probabilities of the consequent $\mathbf{C}^\star_j$ conditioned to antecedent $\mathbf{A}^\star_i$.

Under the interpretation of fuzzy sets in terms of coherent conditional probabilities as stated in equation (1), given any (finitely additive) probability π on an algebra of χ for which $\mu_{A^\star}$ satisfies some suitable restrictions of measurability and setting

$$P(E_{A^\star}) = \int \mu_{A^\star} \mathrm{d}\pi, \tag{3}$$

where the integral is an abstract Stieltjes integral, the assessment $\{\mu_{A^\star}, \pi, P\}$ is coherent. In particular, if π is countably additive and defined on a σ-algebra, the definition above exactly coincides with the "probability of a fuzzy event" as introduced by Zadeh [19].

In our context, assuming the joint probability distribution for the random vector of attributes $\mathbf{X}$ to be expressed by the distribution function $F(\mathbf{x})$, provided $P(E_{\mathbf{A}^\star_i}) > 0$,

we get

$$w_{j|i} = P(E_{\mathbf{C}_j^\star}|E_{\mathbf{A}_i^\star}) = \frac{P(E_{\mathbf{C}_j^\star} \wedge E_{\mathbf{A}_i^\star})}{P(E_{\mathbf{A}_i^\star})} = \frac{\int_{\chi_1 \times \ldots \times \chi_a} \mu_{\mathbf{C}_j^\star}(\mathbf{x}) \odot_t \mu_{\mathbf{A}_i^\star}(\mathbf{x}) \mathrm{d}F(\mathbf{x})}{\int_{\chi_1 \times \ldots \times \chi_a} \mu_{\mathbf{A}_i^\star}(\mathbf{x}) \mathrm{d}F(\mathbf{x})} \quad (4)$$

where $\odot_t$ is any t-norm that guarantees overall coherence. Concerning this last point (see [5]), coherence is guaranteed only by adopting a t-norm in the Frank's class:

$$x \odot_t y = \begin{cases} \min\{x, y\} & \text{if } t = 0 \\ xy & \text{if } t = 1 \\ \max\{0, x + y - 1\} & \text{if } t = +\infty \\ \log_t \left(1 + \frac{(t^x - 1)(t^y - 1)}{(t-1)}\right) & \text{otherwise} \end{cases} \qquad t \in [0, +\infty]. \quad (5)$$

The choice of the parameter t is a degree of freedom in the system so a sensitivity analysis by varying t should be performed.

3 Rule base construction

The construction of rules like (2) is the core of the classifier, and the crucial point is the choice of the fuzzy sets for the antecedents and for the consequents. Once they have been built, conditional probabilities (4) can be estimated either directly by Monte Carlo techniques or through an estimation of the joint probability distribution $F(\mathbf{x})$.

In literature several techniques, like genetic algorithms, fuzzy-neuro systems, parametric estimations and so on, for elicitation of the membership functions $\mu_{\mathbf{A}_i^\star}(\mathbf{x})$ and $\mu_{\mathbf{C}_j^\star}(\mathbf{x})$ have been proposed. We chose a fully non parametric approach, relying on a supervised learning procedure carried on a training set

$$\mathcal{TS} = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\}, \quad (6)$$

where the j-th instance in the training set has attribute description $\mathbf{x}_j$ and a class c_j. Classes c_j are assumed to be linguistic labels belonging to a finite set $\mathcal{C} = \{\mathbf{C}_1^*, \ldots, \mathbf{C}_d^*\}$ and attached by a field expert: for each instance in $\mathcal{TS}$, the expert chooses the label with "highest degree of membership" according to him. Hence, class labels correspond to fuzzy sets whose membership needs to be estimated. For the moment, we limit ourselves to consider training sets with crisp descriptions for all the attributes, hence $\mathbf{x}_i \in \chi_1 \times \ldots \times \chi_a$, $i = 1, \ldots, n$.

Concerning the elicitation of the antecedents membership functions $\mu_{\mathbf{A}_i^\star}(\mathbf{x})$, we chose to estimate marginal membership functions

$$\mu_{A_{k.c}^\star}(x_k), \quad k = 1, \ldots, a, \ c = 1, \ldots, d, \quad (7)$$

defined on each attribute range χ_k, for each consequent label $\mathbf{C}_c^*$, and then to aggregate them through a proper t-norm in the Frank's class:

$$\mu_{\mathbf{A}_i^\star}(\mathbf{x}) = \mu_{A_{1.c_{i1}}^\star}(x_1) \odot_t \ldots \odot_t \mu_{A_{a.c_{ia}}^\star}(x_a), \quad c_{ij} \in \{1, \ldots, d\}, \ j = 1, \ldots, a. \quad (8)$$

Note that the number of potential antecedents is d^a, an hence the number of different rules is $r \leq d^a$ since for some combination of marginal membership functions in (8) the resulting membership can be identically null: this happens whenever at least one membership $\mu_{A^\star_{k.c_{ij}}}(x_k)$ has a disjoint support with respect to some other marginal membership of a different attribute. We neglect antecedents with null membership because we want to base our system only on the training set and in such configurations our sample does not give us any valuable information. The deepening of the admissibility also of antecedents with null memberships is demanded to a future contribution.

Note that, for a fixed attribute range χ_k, we have a marginal membership $\mu_{A^\star_{k.c}}(x_k)$ for each consequent label $\mathbf{C}^\star_c$ because, since the interpretation of the membership function as

$$\mu_{A^\star_{k.c}}(x_k) = P(E_{A^\star_{k.c}} | X_k = x_k), \tag{9}$$

it can be directly estimated through conditional relative frequencies on $\mathcal{TS}$, and the only discriminant information that can be automatically used are the consequent labels. Hence, we will have a membership for each consequent label on each attribute variable domain, with a total number of $a \cdot d$ membership functions.

A "rough" but direct way for estimating (7) is by discretizing - if not yet finite - the range χ_k in a finite number of cells $h_{k1}, \ldots, h_{kn_k}$ (if the r.v. X_k is real such cells are intervals) and for each class to count the relative frequencies:

$$\widetilde{\mu}_{A^\star_{k.c}}(x_k) = \frac{\sharp\{(\mathbf{x}_t, c_t) \in \mathcal{TS} \,|\, (\mathbf{x}_t)_k \in h_{ki} \text{ and } c_t = \mathbf{C}^*_c\}}{\sharp\{(\mathbf{x}_t, c_t) \in \mathcal{TS} \,|\, (\mathbf{x}_t)_k \in h_{ki}\}} \qquad \forall x_k \in h_{ki}, \tag{10}$$

where $(\mathbf{x}_t)_k$ denotes the k-th component of $\mathbf{x}_t$ with $(\mathbf{x}_t, c_t) \in \mathcal{TS}$.

These estimations lead to stepwise membership functions that, by construction, for each dimension $k \in \{1, \ldots, a\}$ form a *proper* (or *strong*) fuzzy partition:

$$\sum_{c=1}^{d} \mu_{A^\star_{k.c}}(x_k) = 1 \qquad \forall x_k \in \chi_k. \tag{11}$$

Such membership functions can be smoothed with different techniques. A simple method that we propose is to replace the crisp partitions $\mathcal{H}_k = \{h_{k1}, \ldots, h_{kn_k}\}$, $k = 1, \ldots, a$, with proper fuzzy grid partitions $\mathfrak{H}_k = \{H^\star_{k1}, \ldots, H^\star_{kn_k}\}$ and consequently approximate the membership functions (7) for $x_k \in \chi_k$ by

$$\widehat{\mu}_{A^\star_{k.c}}(x_k) = \sum_{i=1}^{k_n} \left(\frac{\sum_{\{(\mathbf{x}_t, c_t) \in \mathcal{TS} \text{ and } c_t = \mathbf{C}^\star_c\}} \mu_{H^\star_{k_i}}((\mathbf{x}_t)_k)}{\sum_{\{(\mathbf{x}_t, c_t) \in \mathcal{TS}\}} \mu_{H^\star_{k_i}}((\mathbf{x}_t)_k)} \right) \mu_{H^\star_{ki}}(x_k). \tag{12}$$

Similar considerations can be done to estimate the consequents' membership functions $\mu_{\mathbf{C}^\star_j}(\mathbf{x})$, but now by partitioning the range $\chi_1 \times \ldots \times \chi_a$ either by the crisp partition $\mathbf{H} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_a$ or by the proper fuzzy partition $\mathfrak{H} = \mathfrak{H}_1 \times \ldots \times \mathfrak{H}_a$ and setting

for $\mathbf{x} \in \chi_1 \times \ldots \times \chi_a$

$$\widetilde{\mu}_{\mathbf{C}_j^\star}(\mathbf{x}) = \frac{\sharp\{\mathbf{x}_t \in \mathcal{TS} | \mathbf{x}_t \in \mathbf{h} \text{ and } c_t = \mathbf{C}_j^\star\}}{\sharp\{\mathbf{x}_t \in \mathcal{TS} | \mathbf{x}_t \in \mathbf{h}\}} \qquad \forall \mathbf{x} \in \mathbf{h} \in \mathbf{H}, \tag{13}$$

$$\widehat{\mu}_{\mathbf{C}_j^\star}(\mathbf{x}) = \sum_{\mathbf{K}^\star \in \mathfrak{H}} \left(\frac{\sum_{\{(\mathbf{x}_t, c_t) \in \mathcal{TS} \text{ and } c_t = \mathbf{C}_j^\star\}} \mu_{\mathbf{K}^\star}(\mathbf{x}_t)}{\sum_{\{(\mathbf{x}_t, c_t) \in \mathcal{TS}\}} \mu_{\mathbf{K}^\star}(\mathbf{x}_t)} \right) \mu_{\mathbf{K}^\star}(\mathbf{x}). \tag{14}$$

If membership estimations through frequencies (10) and (13) are used, also the joint probability distribution $F(\mathbf{x})$ can be approximated by the relative frequencies of the training set on $\mathbf{H}$ so that the probabilities of fuzzy events $P(E_{\mathbf{C}_j^\star} \wedge E_{\mathbf{A}_i^\star})$ and $P(E_{\mathbf{A}_i^\star})$ involved in (4) can be approximated by setting

$$p_{\mathbf{h}} = \frac{\sharp\{\mathbf{x}_t \in \mathcal{TS} | \mathbf{x}_t \in \mathbf{h}\}}{\sharp\{\mathbf{x}_t \in \mathcal{TS}\}} \qquad \forall \mathbf{h} \in \mathbf{H}, \tag{15}$$

and then

$$\widetilde{P}(E_{\mathbf{C}_j^\star} \wedge E_{\mathbf{A}_i^\star}) = \sum_{\mathbf{h} \in \mathbf{H}} \left(\widetilde{\mu}_{\mathbf{C}_j^\star}(\mathbf{x}) \odot_t \widetilde{\mu}_{A^\star_{1.c_j}}(\mathbf{x})_1 \odot_t \ldots \odot_t \widetilde{\mu}_{A^\star_{a.c_j}}(\mathbf{x})_a \cdot p_{\mathbf{h}} \right) \tag{16}$$

$$\widetilde{P}(E^\star_{\mathbf{A}_i}) = \sum_{\mathbf{h} \in \mathbf{H}} \left(\widetilde{\mu}_{A^\star_{1.c_j}}(\mathbf{x})_1 \odot_t \ldots \odot_t \widetilde{\mu}_{A^\star_{a.c_j}}(\mathbf{x})_a \cdot p_{\mathbf{h}} \right), \tag{17}$$

respectively, and consequently weights (4) become

$$\widetilde{w}_{j|i} = \frac{\widetilde{P}(E_{\mathbf{C}_j^\star} \wedge E_{\mathbf{A}_i^\star})}{\widetilde{P}(E_{\mathbf{A}_i^\star})}. \tag{18}$$

On the other hand, whenever approximations (12) and (14) are adopted, probabilities of fuzzy events in them can be approximated through Monte Carlo techniques and, consequently, estimates of the conditional probabilities (4) directly computed as

$$\widehat{w}_{j|i} = \widehat{P}(E_{\mathbf{C}_j^\star} | E_{\mathbf{A}_i^\star}) = \frac{\sum_{\mathbf{x} \in \mathcal{TS}} \left(\widehat{\mu}_{\mathbf{C}_j^\star}(\mathbf{x}) \odot_t \widehat{\mu}_{A^\star_{1.c_j}}(\mathbf{x})_1 \odot_t \ldots \odot_t \widehat{\mu}_{A^\star_{a.c_j}}(\mathbf{x})_a \right)}{\sum_{\mathbf{x} \in \mathcal{TS}} \left(\widehat{\mu}_{A^\star_{1.c_j}}(\mathbf{x})_1 \odot_t \ldots \odot_t \widehat{\mu}_{a.A^\star_{c_j}}(\mathbf{x})_a \right)}. \tag{19}$$

Note that the latter quantity does not require estimation of the joint probability distribution $F(\mathbf{x})$. All the conditional probabilities introduced so far can be easily shown to be coherent.

4 Reasoning

Once probabilities of consequents given the antecedents $w_{j|i}$ are estimated either through $\widetilde{w}_{j|i}$ or $\widehat{w}_{j|i}$ as in (18) or (19), it is possible to classify a new case. For the sake of simplicity, in the following we denote with a bullet $^\bullet$ estimates computed either through relative frequencies or their smoothed versions, hence, e.g., $w^\bullet_{j|i}$ stands either for $\widetilde{w}_{j|i}$ or $\widehat{w}_{j|i}$, indifferently.

First of all we have to distinguish if the new case has a crisp or fuzzy description, as detailed in the following two subsections.

4.1 Classification with crisp inputs

In the former case, whenever we have $\mathbf{x}_{\text{new}} \in \chi_1 \times \ldots \times \chi_a$, we can compute the activations $\mu^{\bullet}{}_{\mathbf{A}_i^\star}(\mathbf{x})$, $i = 1, \ldots, r$, of the various rules (2) that represent the "degree of fitness" of the new case with respect to the various antecedents.

Two approaches are now possible. The first one is inspired to the Wang-Mendel generalization given in [8], according to which we foremost select the rules with maximum activation

$$I = \arg \max_{i=1,\ldots,r} \mu^{\bullet}{}_{\mathbf{A}_i^\star}(\mathbf{x}_{\text{new}}), \tag{20}$$

and subsequently we classify through the maximum weight

$$\mathbf{C}^\star_{\text{new}} = \arg \max_{j=1,\ldots,d} \{w^{\bullet}{}_{j|i} | i \in I\}. \tag{21}$$

In general the solution of the above maximization could not be unique and in such a case, either the new instance is not classified since uncertainty remains on the final label, or an imprecise classification with more then one plausible label is allowed.

The second approach is inspired to the Takagi-Sugeno reasoning generalization given in [18]. In this case we want that all the rules contribute to the classification. Fixing a consequent $\mathbf{C}_j^\star$, for each rule we can combine its weight with its activation and, by assuming the conditional independence of claimed consequent $E_{\mathbf{C}_j^\star}$ from the observation $(\mathbf{X} = \mathbf{x}_{\text{new}})$ given the claimed antecedent $E_{\mathbf{A}_i^\star}$, we obtain

$$\begin{aligned} w^{\bullet}{}_{j|i} \cdot \mu^{\bullet}{}_{\mathbf{A}_i^\star}(\mathbf{x}_{\text{new}}) &= P^{\bullet}(E_{\mathbf{C}_j^\star} | E_{\mathbf{A}_i^\star}) \cdot P^{\bullet}(E_{\mathbf{A}_i^\star} | \mathbf{X} = \mathbf{x}_{\text{new}}) \\ &= P^{\bullet}(E_{\mathbf{C}_j^\star} \wedge E_{\mathbf{A}_i^\star} | \mathbf{X} = \mathbf{x}_{\text{new}}). \end{aligned} \tag{22}$$

The conditional independence assumption adopted to obtain (22) is quite natural in our context, since the construction of the fuzzy rule based classifier is founded on the idea that rules are fired by the antecedents, so, once a specific antecedent is claimed the specific value observed $\mathbf{x}_{\text{new}}$ does not matter.

Assuming the almost sure exhaustiveness of the r antecedents built according to Section 3, the disjunction of all combinations like (22) covers all possible combinations of the fixed claimed consequent $E_{\mathbf{C}_j^\star}$ and rules, so that the probabilities

$$P^{\bullet}\left(\bigvee_{i=1}^{a} (E_{\mathbf{C}_j^\star} \wedge E_{\mathbf{A}_i^\star}) \,\middle|\, \mathbf{X} = \mathbf{x}_{\text{new}} \right) \quad j = 1, \ldots, d \tag{23}$$

can be use to classify the new case. Let us stress that (23) is not uniquely determined by the membership functions and the distribution on $\mathbf{X}$: in general, coherence implies that the possible values form a closed interval. Hence, lower $\underline{p}_{j|\mathbf{x}_{\text{new}}}$ and upper bounds $\overline{p}_{j|\mathbf{x}_{\text{new}}}$, $j = 1, \ldots, d$, for (23) can be assessed by coherent extension - e.g., through automatic procedures like that presented in [2] - of the whole conditional probability assessment

$$\left\{ P^{\bullet}(E_{\mathbf{C}_j^\star} \wedge E_{\mathbf{A}_i^\star} | \mathbf{X} = \mathbf{x}_{\text{new}}), \mu^{\bullet}{}_{A^\star_{k.c_j}}(\mathbf{x}_{\text{new}})_k, \mu^{\bullet}{}_{\mathbf{C}_j^\star}(\mathbf{x}_{\text{new}}) \right\}, \tag{24}$$

$j = 1, \ldots, d$, $i = 1, \ldots, r$, $k = 1, \ldots, a$.

The d intervals of coherent values $[\underline{p}_{j|\mathbf{x}_{\text{new}}}, \overline{p}_{j|\mathbf{x}_{\text{new}}}]$ do not determine, in general, a dominant consequent label $\mathbf{C}^{\star}_{\text{new}}$. Hence, also in these cases we can just discard the surely dominated labels, obtaining an imprecise classification.

4.2 Classification with fuzzy inputs

In the second input possibility, whenever we have $\mathbf{B}^{\star}_{\text{new}} \in \mathfrak{F}_{\chi_1 \times \ldots \times \chi_a}$, we cannot compute exact activation values $\mu^*{}_{\mathbf{A}^{\star}_i}(\mathbf{x})$ but we can resort to some similarity grading $\preceq$, e.g., like those proposed in [1], select the group of rules with antecedents $\mathbf{A}^{\star}_i$ more similar to the input $\mathbf{B}^{\star}_{\text{new}}$

$$I^s = \{i \in \{1, \ldots, r\} |\ \nexists l \in \{1, \ldots, r\} \text{ s.t. } (\mathbf{A}^{\star}_i, \mathbf{B}^{\star}_{\text{new}}) \prec (\mathbf{A}^{\star}_l, \mathbf{B}^{\star}_{\text{new}})\}, \tag{25}$$

and then proceed with classification through weight maximization like in (21).

Note that with this kind of input the rule aggregation in Takagi-Sugeno style as done in (22) or (23) is not possible since we cannot use activation levels.

5 Empirical study

The final goal of the present study is to implement the whole classification method described in previous sections in `R` software [15].

Due to space limitations we show only some preliminary results focusing on datasets with crisp attribute descriptions, both in the training set and in the test set, and membership estimations performed through frequencies approximations. We first consider the following example in which our method is applied to a doping alert problem.

Example 1 *A medical center has collected* 2000 *profiles of non-professional athletes, the half of which have been claimed to use illegal doping drugs. The profiles are described by means the variables BMI = "percentage of body mass increment" and PVC = "percentage of packed cell volume". Figure 1 shows the plot of the collected dataset together with the corresponding class "Doped" (in red) and "Undoped" (in yellow).*

We apply our method to the doping dataset executing a 5-fold cross-validation, where in each fold we select 1600 elements, 800 from each class, to form the training set $\mathcal{TS}$, leaving the remaining 400 as test set. Table 1 shows average percentage of correctly classified instances varying the parameter t of the Frank's t-norm used to realize fuzzy operations. Results contained in Table 1 highlight that the proposed algorithm has a good robustness with respect to the choice of the parameter t, reaching the best performance for $t = e$.

In the rest of this section, the parameter t of the t-norm is fixed to $t = e$ since, as shown in Example 1, the algorithm presents a good robustness with respect to the choice of the t-norm to realize fuzzy operations. Moreover, the best results are obtained for $t = e$.

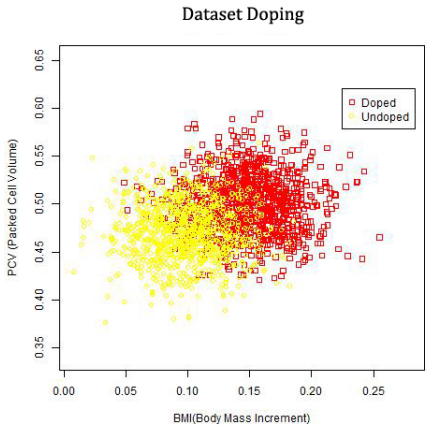


Figure 1: Plot of the doping dataset

t	Avg. CCI
0	$80.20\% \pm 1.77\%$
1	$79.85\% \pm 2.24\%$
$+\infty$	$80.25\% \pm 1.25\%$
e	$81.60\% \pm 1.48\%$

Table 1: Sensitivity analysis for t

We compare our probabilistic fuzzy rule based classifier with Wang-Mendel classifier and with the one discussed in [8]. At this aim we randomly generate four datasets (namely A, B, C and D) of 3000 examples, divided in three groups of size 1000, each for a different class label. For each class label, the 1000 examples are generated with a bivariate normal distribution with mean $\bar{\mu}$, whose marginal distributions are independent and with the same standard deviation σ, the latter being fixed for the three classes. The characteristics of the generated datasets are reported in Table 2, while the plots of datasets A and C are reported in Figure 2. In [8] also a fifth dataset E is generated: we decided not to report it since it is not particularly meaningful in a comparison test, as the classes result to be practically indistinguishable.

Dataset	$\bar{\mu}$ of class 1	$\bar{\mu}$ of class 2	$\bar{\mu}$ of class 3	σ
A	$(10, 10)$	$(10, 30)$	$(30, 10)$	9
B	$(10, 10)$	$(10, 20)$	$(20, 10)$	9
C	$(10, 10)$	$(10, 30)$	$(30, 10)$	12
D	$(10, 10)$	$(10, 20)$	$(20, 10)$	12

Table 2: Structure of the generated datasets

For each dataset, we execute a 5-fold cross-validation. So, in each fold we select 2400 elements, 800 from each class, to form the training set $\mathcal{TS}$, leaving the remaining 600 as test set.

The following Table 3 lists the average percentage of correctly classified instances of our method (CPP), compared with the performance of the Wang-Mendel algorithm (WM) and that of the de Melo-Lucas-Delgado algorithm (MLD).

According to Table 3, CPP always performs better than WM but its performance is slightly worse than that of MLD. We conjecture that the better performance of MLD algorithm is essentially determined by the particular choice of membership functions

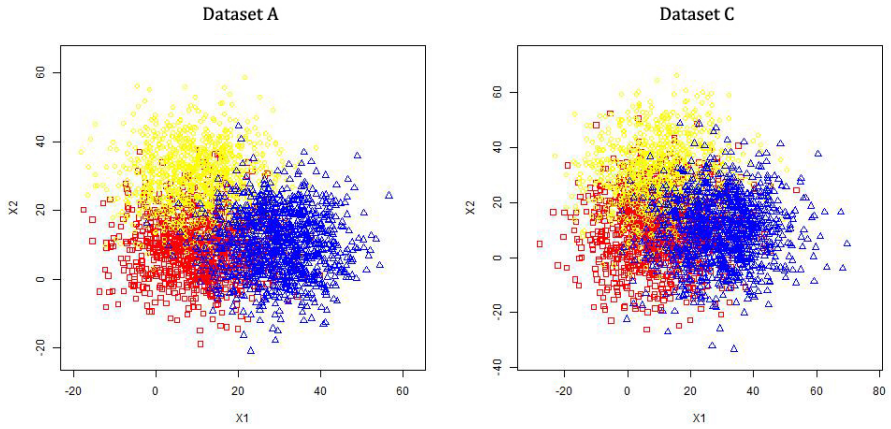


Figure 2: Plots of datasets A and C

Dataset	Alg.	Avg. CCI	Dataset	Alg.	Avg. CCI
A	WM	$62.13\% \pm 13.38\%$	B	WM	$47.97\% \pm 7.55\%$
	MLD	$81.10\% \pm 4.59\%$		MLD	$67.00\% \pm 2.26\%$
	CPP	$\mathbf{77.33}\% \pm \mathbf{1.61}\%$		**CPP**	$\mathbf{57.87}\% \pm \mathbf{1.73}\%$
C	WM	$59.87\% \pm 6.88\%$	D	WM	$34.97\% \pm 9.07\%$
	MLD	$69.13\% \pm 5.58\%$		MLD	$60.00\% \pm 3.03\%$
	CPP	$\mathbf{65.40}\% \pm \mathbf{0.28}\%$		**CPP**	$\mathbf{52.37}\% \pm \mathbf{0.47}\%$

Table 3: Performance of algorithms WD, MLD and CPP

which is done "a priori": in such a method, indeed, Gaussian membership functions are chosen and their choice seems to be particularly fitted to the given datasets. Our algorithm does not require a preliminary assessment of membership functions, thus is evidently more general and versatile.

6 Conclusion

This paper is a first contribution to design a probabilistic fuzzy rule based classifier in the setting of coherent conditional probabilities. We feel that the proposal could be of help when applied to real classification problems, as that one of the doping alert, since the determination of the rules gives an immediate interpretability of the classification mechanism. Moreover, as the consequents classification labels are fuzzy and endowed with probabilities evaluations, a richer information is available with respect to purely fuzzy classifiers.

Due to space limitations, we only provided a sketch of the proposed algorithm together with some preliminary experimental results and comparisons. Several aspects

must be deepened, both from a theoretical and a experimental point of view, and they will be the subject of future publications.

In particular, the following points have to be considered: estimation techniques of the various membership functions must be refined; aggregation of the different rules should be enriched with a stochastic mechanism that could be derived by probabilistic properties of the various computed conditional probabilities of fuzzy sets; a systematic sensitivity analysis with respect to the choice of the t-norm parameter is needed, and a complete empirical study, both on artificial and real data, should be carried on.

Acknowledgments

This work has been supported by the Italian Ministry of Health under grant `J52I14001640001` *"Sistemi intelligenti di ausilio alle decisioni per l'identificazione precoce e la dissuasione all'utilizzo del doping"*.

References

[1] Bouchon-Meunier, B., Coletti, G., Lesot, M.-J., Rifqi, M.: Towards a conscious choice of a fuzzy similarity measure: a qualitative point of view. In E. Hüllermeier et al (eds.), *Comp. Int, for Know.-B. Sys. Des.*, Lec. Notes in Comp. Sci. Volume 6178, pp 1-10 (2010)

[2] Capotorti, A., Vantaggi, B.: Locally strong coherence in an inference process. *Ann. Math. Artif. Int.*, 35(14), pp 125-149 (2002)

[3] Coletti, G., Petturiti, D., Vantaggi, B.: Probabilistic Fuzzy Reasoning in a Coherent Setting. In *8th Con. of the Eur. Soc. for Fuzzy Log. and Tech. (EUSFLAT 2013)* (2013)

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

[5] Coletti, G., Scozzafava, R.: Conditional probability, fuzzy sets, and possibility: a unifying view, *Fuzzy Sets and Sys.*, 144, pp 227–249 (2004)

[6] Coletti, G., Scozzafava, R.: Conditional probability and fuzzy information, *Computational Statistics & Data Analysis*, 51, pp 115–132 (2006)

[7] Coletti, G., Vantaggi, B.: Probabilistic Reasoning in a Fuzzy Context. In L.A. Zadeh et al (eds.), *Rec. Dev. and New Dir. in Soft Comp.*, Stud. in Fuzz. and Soft Comp. Volume 317, pp 97–115 (2014)

[8] de Melo, L.-G., Lucas, L.-A., Delgado, M.-R.: Rule-Base Design Using Probabilistic Weights: A Preliminary Analysis of Uncertainty Aspects. In S. Greco et al (eds.), *Adv. in Comp. Int.*, Comm. in Comp. and Inf. Sci.Volume 300, pp 655–664 (2012)

[9] Frank, M.J.: On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequat. Math.*, 19, pp 194–226 (1979)

[10] Gómez-Skarmeta, A., Jiménez, F.: Fuzzy modeling with hybrid systems. *Fuzzy Sets and Sys.*, 104(2), pp 199-208 (1999)

[11] Liu, Z., Li, H.-X.: A Probabilistic Fuzzy Logic System for Modeling and Control. *IEEE Trans. on Fuzzy Sys.*, 13, pp 848–859 (2005)

[12] Meghdadi, A.H., Akbarzadeh-T, M-R: Probabilistic Fuzzy Logic and Probabilistic Fuzzy Systems. In *10th IEEE Int. Conf. on Fuzzy Sys.*, pp 1127–1130 (2001)

[13] Meghdadi, A.H., Akbarzadeh-T, M-R: Uncertainty Modeling through Probabilistic Fuzzy Systems. In *Proc. of the 4th Int. Symp. on Unc. Model. and Anal. (ISUMA 2003)* (2003)

[14] Parandehgheibi, M.: Probabilistic Classification using Fuzzy Support Vector Machines. In *Proc. of the 6th INFORMS Work. on Data Min. and Health Inf. (DM-HI 2011)* (2011)

[15] R Development Core Team: R a language and environment for statistical computing. R Foundation for Statistical Computing, `http://www.R-project.org`.

[16] van den Berg, J., Kaymak, U., Almeida, R.J.: Conditional Density Estimation Using Probabilistic Fuzzy Systems. *IEEE Trans. on Fuzzy Sys.*, 21(5), 869–882 (2013)

[17] van den Berg, J., van den Bergh, W.-M., Kaymak, U.: Probabilistic and statistical fuzzy set foundations of competitive exception learning. In *Proc. of 10th IEEE Int. Conf. on Fuzzy Sys.*, 2, 1035–1038 (2001)

[18] Waltman, L., Kaymak, U., van den Berg, J.: Maximum likelihood parameter estimation in probabilistic fuzzy classifiers. In *Proc. 14th IEEE Int. Conf. Fuzzy Sys.*, 1098–1103 (2005)

[19] Zadeh, L.A.: Probability measures of Fuzzy events. *J. of Math. Anal. and App.*, 23(2), 421–427 (1968)

Fuzzy Sets through Likelihood in Probabilistic and Possibilistic Frameworks

Giulianella Coletti
Dept. Matematica e Informatica
University of Perugia
giulianella.coletti@unipg.it

Davide Petturiti
Dept. Matematica e Informatica
University of Perugia
davide.petturiti@unipg.it

Barbara Vantaggi
Dept. S.B.A.I.
"La Sapienza" University of Rome
barbara.vantaggi@sbai.uniroma1.it

Abstract

We propose an interpretation of fuzzy membership as coherent T-conditional possibility (where T is a continuous triangular norm) regarded as a function of the conditioning event (possibilistic likelihood) and we check which operations between fuzzy sets arise in this framework. Since in the literature an interpretation of the membership as probabilistic likelihood has already been studied, we provide a comparison between the possibilistic and the probabilistic frameworks.

1 Introduction

Models and tools for jointly handling uncertainty and vagueness have to be performed in order to deal with problems involving both these aspects. In fact, incomplete, linguistic and vague information can coexist in real problems, mainly due to the presence of several heterogeneous sources of knowledge. This fact generates new problems in probability and statistics, and so many methods and techniques have been proposed in literature, which combine probability, statistics and fuzzy methods.

Notice that in order to combine uncertain information with vagueness, we need to refer to notions of coherence, which guarantee an effective tool for controlling global consistency and ruling the inferential procedures.

We recall that a general inferential problem can be simply seen as an extension of an uncertainty measure assessment to other events, maintaining consistency with the framework of reference.

In the probabilistic framework this problem reduces in the simplest case, i.e., when the likelihood and the prior distribution are completely assessed on the same finite

space, to Bayes rule. However, the hypotheses of Bayes theorem are usually not satisfied and so we need to handle generalized Bayesian inferential procedures whose result is in general not unique, but consists in an interval of coherent values for any conditional event.

Aim of this paper is to merge in the context of possibility theory both prior uncertainty quantified by a possibility measure (for instance obtained as upper envelope of the extensions of a probability related to the range of a different variable [7]) and the vagueness expressed by fuzzy sets.

The combination of probabilistic uncertainty and vagueness is possible thanks to the interpretation, given by Coletti and Scozzafava [5, 6], of the fuzzy membership as a coherent conditional probability. In [11], an analogous interpretation in terms of coherent lower and upper conditional probabilities has been given.

Now, following the same line, an interpretation of membership function as coherent T-conditional possibility (where T is a continuous triangular norm) is provided. The semantic behind this new interpretation follows the one of [5, 6, 11]: for every x in the range of a variable X, the value of the membership $\mu_\varphi(x)$ of a fuzzy set related to a property φ of X is *the measure of how much You believe in the Boolean event "You claim that X has property φ" when $X = x$.* One of the main reason for interpreting the membership as a possibilistic likelihood resides in the semantic meaning of membership (see [13]).

First of all we check which operations between fuzzy sets arise under the possibilistic interpretation and then a comparison between possibilistic and probabilistic memberships and their operations is drawn.

As discussed before, our main aim is to make inference starting from a possibility measure on the algebra spanned by the events $\{X = x\}$ and a family of possibilistic likelihoods (the membership functions). In the possibilistic setting, to handle this procedure it is necessary, first of all, to check whether possibilistic and fuzzy information are globally coherent with respect to the chosen definition of conditioning. Then the general problem of checking coherence and that of making inference need to be deepened, in order to find relevant results. Thus, we analyse some updating rules and we discuss their implications in the inferential procedures, also providing a comparison.

Furthermore we show how to create a collection of possibilistic fuzzy IF-THEN rules: in particular we discuss how to compute the possibility associated to these rules and how to propagate coherent intervals, when either the premise or the consequence is fuzzy. The latter point relies on a notion of inclusion given in [17].

2 Coherent T-conditional possibilities

We refer to the notion of T-conditional possibility (with T a continuous t-norm) introduced in [1, 10]. Coherence is well-known in probability theory starting from the famous characterization given by de Finetti [12]. The same notion has been studied also in other frameworks and in particular in possibility theory (see [10]) by referring to the axiomatic definition of T-conditional possibility.

Theorem 1 *Let T be a continuous t-norm, $\mathcal{G} = \{E_1|H_1, \ldots, E_n|H_n\}$ an arbitrary set of conditional events, and $\mathcal{C}_0$ and $\mathcal{B}$ the set of atoms and the algebra spanned by $\{E_1, H_1, \ldots, E_n, H_n\}$, respectively.*
For a real function $\Pi : \mathcal{G} \to [0,1]$, the following statements are equivalent:

a) Π is a coherent T-conditional possibility assessment on $\mathcal{G}$;

b) there exists a sequence of compatible systems $\mathcal{S}_\alpha^\Pi$ $(\alpha = 0, \ldots, k)$, with unknowns $x_r^\alpha \geq 0$ for $C_r \in \mathcal{C}_\alpha$,

$$\mathcal{S}_\alpha^\Pi = \begin{cases} \max\limits_{C_r \subseteq E_i \wedge H_i} x_r^\alpha = T\left(\Pi(E_i|H_i), \max\limits_{C_r \subseteq H_i} x_r^\alpha\right) \\ \left[\text{for all } E_i|H_i \in \mathcal{G} \text{ such that } \max\limits_{C_r \subseteq H_i} \xi_r^{\alpha-1} < 1\right] \\ \xi_r^{\alpha-1} = T\left(x_r^\alpha, \max\limits_{C_j \in \mathcal{C}_\alpha} \xi_j^{\alpha-1}\right) \qquad \text{if } C_r \in \mathcal{C}_\alpha \\ \max\limits_{C_r \in \mathcal{C}_\alpha} x_r = 1 \end{cases}$$

with $\alpha = 0, \ldots, k$, where $\bar{\xi}^\alpha$ (with r-th component ξ_r^α) is the solution of $\mathcal{S}_\alpha^\Pi$ and $\mathcal{C}_\alpha = \{C_r \in \mathcal{C}_{\alpha-1} : \xi_r^{\alpha-1} < 1\}$, moreover $\xi_r^{-1} = 0$ for any $C_r \in \mathcal{C}_0$.

We recall that any unconditional possibility $\Pi(\cdot)$ can be considered as a coherent T-conditional possibility setting $\Pi(\cdot) = \Pi(\cdot|\Omega)$, where Ω denotes the sure event. Moreover, for $T = \min$ or a strict t-norm, every coherent T-conditional possibility assessment can be extended to any new conditional event and the coherent extension lays in a closed interval (see [10]).

2.1 Possibilistic likelihood

We recall some results proved in [4]:

Theorem 2 *Let $\mathcal{L} = \{H_i\}_{i=1,\ldots,n}$ be a finite partition of Ω and E an event. For every function $f : \{E\} \times \mathcal{L} \to [0,1]$ satisfying condition*

(L1) $f(E|H_i) = 0$ if $E \wedge H_i = \emptyset$ and $f(E|H_i) = 1$ if $H_i \subseteq E$

the following statements hold:

i) f is a coherent conditional probability;

ii) f is a coherent T-conditional possibility (for every t-norm T).

The above result shows a common property between probabilistic and possibilistic (point) likelihood, so this allows to regard a probabilistic likelihood as a possibilistic one and vice versa.

Moreover, it emphasizes that no significant property characterizes likelihood as point function, so in the sequel we call *likelihood function* any function $f : \{E\} \times \mathcal{L} \to [0,1]$, with $\mathcal{L} = \{H_i\}_{i=1,\ldots,n}$, a finite partition of Ω, satisfying condition *(L1)*.

We consider now a finite class $\mathcal{F}$ of likelihood functions f_j, each defined on $\{E_j\} \times \mathcal{L}$, for $j = 1, \ldots, m$, and we study global coherence taking into account also a "prior" on $\mathcal{L}$ in both in the possibilistic and the probabilistic frameworks.

In what follows recall that events in a family $\mathcal{E} = \{E_j\}_{j=1,\ldots,m}$ are *logically independent* if $E_1^* \wedge \ldots \wedge E_m^* \neq \emptyset$, where E_j^* stands either for E_j or E_j^c.

Theorem 3 *Let $\mathcal{L} = \{H_i\}_{i=1,\ldots,n}$ be a finite partition of Ω, $\mathcal{E} = \{E_j\}_{j=1,\ldots,m}$ a set of logically independent events, and $\mathcal{F} = \{f_j\}_{j=1,\ldots,m}$ a set of likelihood functions on $\mathcal{E} \times \mathcal{L}$. Let p and π be a probability and a possibility distribution on $\mathcal{L}$. The following conditions hold:*

i) the assessment $\{\mathcal{F}, p\}$ is a coherent conditional probability;

ii) the assessment $\{\mathcal{F}, \pi\}$ is a coherent T-conditional possibility (for every continuous t-norm T).

3 Fuzzy sets as coherent T-conditional possibilities

Following the interpretation of fuzzy sets in terms of coherent conditional probabilities, we give an interpretation based on coherent T-conditional possibilities.

Let X be a (not necessarily numerical) variable, with range $\mathcal{C}_X$, and, for any $x \in \mathcal{C}_X$, let us indicate by x the event $A_x = \{X = x\}$.

Let φ be any *property* related to the variable X and let us refer to the state of information of a real (or fictitious) person that will be denoted by "You".

Let us consider the Boolean event $E_\varphi =$ "*You claim that X has property φ*", then we can give the following definition of fuzzy set E_φ^*:

Definition 1 *Let X be any variable with range $\mathcal{C}_X$, φ a related property and $\Pi(E_\varphi|x)$, for $x \in \mathcal{C}_X$, a coherent T-conditional possibility assessment. A fuzzy subset E_φ^* of $\mathcal{C}_X$ is a pair*

$$E_\varphi^* = \{E_\varphi\, ,\, \mu_\varphi\}, \tag{1}$$

with $\mu_\varphi(x) = \Pi(E_\varphi|x)$ for every $x \in \mathcal{C}_X$.

Then we can interpret the membership function $\Pi(E_\varphi|x)$, for $x \in \mathcal{C}_X$, as the measure (in the possibilistic framework) of Your degree of belief in E_φ, when X assumes the different values of its range.

It follows from [10] that the lower or upper envelope of a set of membership functions related to the same property φ and the same X is still a membership function.

3.1 Operations

By following [5], the binary operations of union and intersection and that of complementation between fuzzy sets, can be directly obtained by using the rules of coherent T-conditional possibility and the logical independence between E_φ and E_ψ. Notice that the events E_φ and E_ψ are "usually" logically independent, in particular they are

logically independent when $\psi = \neg\varphi$: indeed, we can claim both "X has the property φ" and "X has the property $\neg\varphi''$, or only one of them or finally neither of them. Similarly, E_φ and E_ψ are logically independent in case ψ is the superlative of φ.

Let us denote by $\varphi \vee \psi$ and $\varphi \wedge \psi$, respectively, the properties "φ or ψ", "φ and ψ" (note that the symbols $\wedge$ and $\vee$ do not indicate Boolean operations, since φ and ψ are not Boolean objects) and define:

$$E_{\varphi\vee\psi} = E_\varphi \vee E_\psi, \tag{2}$$
$$E_{\varphi\wedge\psi} = E_\varphi \wedge E_\psi. \tag{3}$$

Let us consider two properties φ and ψ related to the same variable X, such that E_φ and E_ψ are logically independent, and let us consider the relevant fuzzy subsets E^*_φ and E^*_ψ on $\mathcal{C}_X$.

For any given x in $\mathcal{C}_X$, the assessment $\Pi(E_\varphi \wedge E_\psi|x) = v$ is coherent if and only if it takes values in the interval

$$0 \leq v \leq \min\{\Pi(E_\varphi|x), \Pi(E_\psi|x)\}. \tag{4}$$

The inequality above puts in evidence a first difference between the probabilistic and the possibilistic interpretations, in fact in the probabilistic setting the two bounds coincide with the Frechet bounds.

In the probabilistic interpretation, fixed the value for the membership function of the fuzzy intersection, the value for the membership function of the fuzzy union is uniquely determined, while in the possibilistic interpretation, independently of the value of $\Pi(E_\varphi \wedge E_\psi|x)$, we get a unique value for the fuzzy union which is

$$\Pi(E_\varphi \vee E_\psi|x) = \max\{\Pi(E_\varphi|x), \Pi(E_\psi|x)\}. \tag{5}$$

This allows to define fuzzy union and intersection as follows:

$$E^*_\varphi \cup E^*_\psi = \{E_{\varphi\vee\psi}, \mu_{\varphi\vee\psi}\}, \tag{6}$$
$$E^*_\varphi \cap E^*_\psi = \{E_{\varphi\wedge\psi}, \mu_{\varphi\wedge\psi}\}, \tag{7}$$

with

$$\mu_{\varphi\vee\psi}(x) = \Pi(E_\varphi \vee E_\psi|x), \tag{8}$$
$$\mu_{\varphi\wedge\psi}(x) = \Pi(E_\varphi \wedge E_\psi|x). \tag{9}$$

Finally, denoting by $(E^*_\varphi)' = E^*_{\neg\varphi} = (E_{\neg\varphi}, \mu_{\neg\varphi})$ the complementary fuzzy set of E^*_φ, we have

$$\mu_{\neg\varphi}(x) = 1 - \mu_\varphi(x) = 1 - \Pi(E_\varphi|x) = \Pi(E_{\neg\varphi}|x), \tag{10}$$

which is a (possibilistic) likelihhod and so a coherent T-conditional possibility.

Notice that also in this framework, the relation $E_{\neg\varphi} \neq (E_\varphi)^c$ holds, in fact, while

$$E_\varphi \vee (E^c_\varphi) = \Omega,$$

due to the logical independence of E_φ and $E_{\neg\varphi}$, we have instead

$$E_\varphi \vee E_{\neg\varphi} \subseteq \Omega,$$

so, if we consider the union of a fuzzy subset and its complement

$$E^*_\varphi \cup (E^*_\varphi)' = \{E_{\varphi\vee\neg\varphi}, \mu_{\varphi\vee\neg\varphi}\}$$

we obtain in general a *fuzzy subsetset* of $\mathcal{C}_X$.

For two fuzzy subsets E^*_φ and E^*_ψ, corresponding to the random quantities X and Y, respectively, the following choice for the membership of conjunction and disjunction is a coherent T-conditional possibility:

$$\mu_{\varphi\vee\psi}(x,y) = \Pi(E_\varphi \vee E_\psi | A_x \wedge A_y), \tag{11}$$
$$\mu_{\varphi\wedge\psi}(x,y) = \Pi(E_\varphi \wedge E_\psi | A_x \wedge A_y), \tag{12}$$

with the only constraints

$$0 \leq \mu_{\varphi\wedge\psi}(x,y) \leq \min\{\mu_\varphi(x), \mu_\psi(y)\}, \tag{13}$$
$$\mu_{\varphi\vee\psi}(x,y) = \max\{\mu_\varphi(x), \mu_\psi(y)\}, \tag{14}$$

under the assumption $\Pi(E_\varphi|A_x \wedge A_y) = \Pi(E_\varphi|A_x)$ and $\Pi(E_\psi|A_x \wedge A_y) = \Pi(E_\psi|A_y)$.

Notice that the last constraint can be interpreted as a conditional independence assumption, which is reasonable in this context.

4 Possibility of "fuzzy events"

For simplicity, we refer to variables X with a finite range. First of all, recall that the concept of fuzzy event, as introduced by Zadeh, is for us an ordinary event of the kind

$$E_\varphi = \text{"You claim that } X \text{ has property } \varphi\text{"}.$$

As stated in Theorem 3, for every (possibilistic) likelihood $\Pi(E|x)$ and for every possibility distribution $\Pi(x)$, the global assessment is coherent. Then by using our interpretation,

$$\{\mu_\varphi(x), \Pi(x)\}_{x\in\mathcal{C}_X}$$

is a coherent T-conditional possibility and so it is coherently extendible to E_φ.

It is easy to see that the only coherent value for the probability of E_φ is

$$\Pi(E_\varphi) = \max_{x\in\mathcal{C}_X} T(\mu_{\varphi_i}(x), \Pi(x)), \tag{15}$$

which is a generalization of Sugeno integral [16], where T is the continuous t-norm used for defining the T-conditional possibility.

This leads to the possibility of a "fuzzy event", which is the counterpart of Zadeh's definition of the probability of a "fuzzy event" [18].

Now let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector with range $\mathcal{C}_{\mathbf{X}}$, where each component X_h has range $\mathcal{C}_{X_h}$. Let $\mathcal{F}(\mathcal{C}_{\mathbf{X}})$ be a finite family of fuzzy subsets $E^*_{\varphi_i} = \{E_{\varphi_i}, \mu_{\varphi_i}\}$, with $i \in I$, related to the (possibly coincident) components X_i of $\mathbf{X}$, where the events $\{E_{\varphi_i}\}_{i\in I}$ are assumed to be logically independent.

For every joint possibility distribution Π on the events $\{\mathbf{X} = \mathbf{x}\} = (X_1 = x_1, ..., X_n = x_n)$, the global assessment $\{\mu_{\varphi_i}, \Pi\}_{i\in I}$ is a coherent T-conditional possibility [4].

Moreover, it is easy to prove that setting

$$\begin{aligned} \Pi(E_{\varphi_i}) &= \max_{x_i \in \mathcal{C}_{X_i}} T(\mu_{\varphi_i}(x_i), \Pi(x_i)), \\ \Pi(E_{\varphi_i} \wedge E_{\varphi_j}) &= \max_{(x_i,x_j)\in \mathcal{C}_{(X_i,X_j)}} T(\mu_{\varphi_i \wedge \varphi_j}(x_i, x_j), \Pi(x_i, x_j)), \end{aligned}$$

for any choice of $\mu_{\varphi_i \wedge \varphi_j}$ in the interval defined by (4), the possibility assessment

$$\{\mu_{\varphi_i}, \Pi\}_{i\in I} \cup \{\Pi(E_{\varphi_i}), \Pi(E_{\varphi_i} \wedge E_{\varphi_j})\}_{i,j\in I}$$

is still coherent. Furthermore, the extension to $E_{\varphi_i} \vee E_{\varphi_j}$ is uniquely determined by (5).

The above assessment Π is a coherent T-conditional possibility, so by using the extension Theorem [10] it can be extended further to any new conditional event $A|B$ where A, B, with $B \neq \emptyset$, are events of the algebra $\mathcal{B}$ spanned by $\{E_{\varphi_i}\}_{i\in I} \cup \{A_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{C}_{\mathbf{X}}}$. This extension is not unique in general and, in particular, for the events $A|B$, with $A = E_{\varphi_i}$ and $B = E_{\varphi_j}$ the coherent extension $\Pi(E_{\varphi_i}|E_{\varphi_j})$ is 1 if $i = j$ and for $i \neq j$ it is a solution of the equation

$$\Pi(E_{\varphi_i} \wedge E_{\varphi_j}) = T(x, \Pi(E_{\varphi_j})). \tag{16}$$

Remark 1 *As in the probabilistic framework, the values $\Pi(E_{\varphi_i}|E_{\varphi_j})$ computed by the formula above are coherent only when the events E_{φ_i} and E_{φ_j} are logically independent, so, for instance the same formula cannot be used for obtaining the coherent extension of Π to $E_{\varphi_i}|E_{\varphi_i}$ which is necessarily 1.*

In the case T is a strict t-norm, equation (16) has not unique solution only in the case $\Pi(E_{\varphi_j}) = 0$. We notice that for a strict t-norm T, one has $\Pi(E_{\varphi_j}) = 0$ if and only if $\Pi(x_j) = 0$ for every $x_j \in \mathcal{C}_{X_j}$ such that $\mu_{\varphi_j}(x_j) > 0$.

In this case to obtain a unique extension we need to specify the T-conditional possibility $\Pi(\cdot|B)$, where B is the logical sum of the events $x_j \in \mathcal{C}_{X_j}$ such that $\mu_{\varphi_j}(x_j) = \Pi(E_{\varphi_j}|x_j) = 0$.

In this case, since $1 = \Pi(B|B) = \max_{x_j \subseteq B} \Pi(x_j|B)$, at least one event $x_j \subseteq B$ is such that $\Pi(x_j|B) = 1$ and so $\Pi(E_{\varphi_i}|E_{\varphi_j})$ is the unique solution (see [10, 2]) of the equation:

$$\Pi(E_{\varphi_i} \wedge E_{\varphi_j}|B) = T(x, \Pi(E_{\varphi_j}|B)),$$

where $\Pi(E_{\varphi_j}|B) = \max_{x_j \in \mathcal{C}_{X_j}} T(\mu_{\varphi_j}(x_j), \Pi(x_j|B))$ and similarly for $\Pi(E_{\varphi_i} \wedge E_{\varphi_j}|B)$.

Analogous considerations hold (mutatis mutandis) in case T is the minimum t-norm.

5 Possibilistic fuzzy reasoning

A significant problem in fuzzy literature is managing fuzzy rule based systems which are essentially composed by a set of IF-THEN rules of the form

"IF A THEN B: with a given degree",

where either the premise A or the consequence B of the rule can be fuzzy sets [14, 15, 17, 19], and the degree could be interpreted as a possibility (or another uncertainty measure) evaluation.

A typical example consists in a possibilistic fuzzy IF-THEN classifier which aims at determining the class of Y in $\{C_1, \ldots, C_t\}$ to which a data point $\mathbf{x} = (x_1, \ldots, x_n)$ belongs. Let us stress that elements of $\{C_1, \ldots, C_t\}$ can be taken either as crisp or fuzzy classes on $\mathcal{C}_Y$. This task can be faced introducing a possibilistic fuzzy IF-THEN classifier formed by a set of rules of the form:

"IF $\mathbf{X}$ IS E_{φ_j} THEN Y IS C_k: with possibility $\pi_{k|j}$".

Notice that the values $\pi_{k|j}$ could be seen as values of degree of inclusion [9].

Example 1 *A well-known Italian factory of vintage scooters launched a new model called Rétro, for which the customer can choose among a fixed number of combinations of engine sizes and color configurations, elaborated after a preliminary market survey. Let us denote with S and M the variables "seat color" and "rear-view mirror color" whose possible values are*

$$\begin{aligned} \mathcal{C}_S &= \{s_1 = \text{"black"}, s_2 = \text{"brown"}, s_3 = \text{"beige"}\}, \\ \mathcal{C}_M &= \{m_1 = \text{"black"}, m_2 = \text{"white"}, m_3 = \text{"metal"}\}. \end{aligned}$$

Thus consider the vector (S, M) whose range is $\mathcal{C}_S \times \mathcal{C}_M$.

Rétro 150cc is produced only with the metal rear-view mirror and not with the black seat, Rétro 125cc is produced not with the black seat, while Rétro 50cc has no restriction on combinations.

The marketing division of the factory singled out the variable M as an indicator of a scooter which is more juvenile or more vintage and aims at determining the impact of variable S on M. Concerning variable S, the properties $l =$ "light" and $d =$ "dark" are considered, while for M the properties $j =$ "juvenile" and $v =$ "vintage" are taken into account together with the following membership functions:

$\mathcal{C}_S$	s_1	s_2	s_3
μ_l	0	0.2	1
μ_d	1	0.8	0

$\mathcal{C}_M$	m_1	m_2	m_3
μ_j	1	0.8	0.2
μ_v	0	0.5	1

The production will start in two months, so a quantification of joint uncertainty on (S, M) is not available at present moment.

Let E be the variable "engine size" whose set of possible values is

$$\mathcal{C}_E = \{e_1 = \text{"50cc"}, e_2 = \text{"125cc"}, e_3 = \text{"150cc"}\}.$$

Since $\mathcal{C}_E$ exhausts the possible engine sizes of all models of scooter produced by the factory, a probabilistic quantification of uncertainty on E can be obtained through the selling frequencies of the last year, obtaining

$$P(e_1) = 60\%, \quad P(e_2) = P(e_3) = 20\%.$$

It is easy to prove that the partition of the sure event determined by (S, M) is weakly logically independent (see [7]) of the partition determined by E, thus the upper probability induced by P on the algebra generated by (S, M) is a possibility measure having the following distribution:

$M \downarrow S \rightarrow$	s_1	s_2	s_3
m_1	0.6	0.8	0.8
m_2	0.6	0.8	0.8
m_3	0.6	1	1

Taking $T = \min$, simple computations allow to determine the following possibilistic fuzzy IF-THEN rule base linking variable S to variable M:

- *IF S IS "light" THEN M IS "juvenile":* $\Pi(E_j|E_l) = 0.8$,
- *IF S IS "light" THEN M IS "vintage":* $\Pi(E_v|E_l) = 1$,
- *IF S IS "dark" THEN M IS "juvenile":* $\Pi(E_j|E_d) \in [0.8, 1]$,
- *IF S IS "dark" THEN M IS "vintage":* $\Pi(E_v|E_d) \in [0.8, 1]$.

More generally, we start from a list of IF-THEN rules (weak implications) forming a set $\mathcal{D}$ of (ordered) pairs (E_φ, E_ψ) of fuzzy subsets, with degree

$$I(E_\psi, E_\varphi) = \Pi(E_\varphi|E_\psi)$$

of fuzzy inclusion (of E_ψ in E_φ) equal to 1, and call any such set a *Maximum Degree of Fuzzy Inclusion set*, or *MDFI-set* for short.

Given $\mathcal{D}$, the problem is to find further pairs of fuzzy subsets in $\mathcal{F}(\mathcal{C}_\mathbf{X})$ with maximum degree of fuzzy inclusion (or *MDFI-pairs*).

Even if for a coherent assessment on an arbitrary set of conditional events $\mathcal{G}$ its enlargement to a family $\mathcal{G}' \supset \mathcal{G}$ is generally not unique, for some events we can have a unique coherent extension which allows to define the important concept of *entailment*.

A MDFI-set $\mathcal{D}$ *entails* the pair (E_φ, E_ψ) of fuzzy sets with degree belonging to an interval $[\pi', \pi'']$ if the coherent value for $\Pi(E_\psi|E_\varphi)$ ranges in $[\pi', \pi'']$. In particular, the MDFI-set $\mathcal{D}$ *strictly entails* the pair (E_φ, E_ψ) if the only coherent value for $\Pi(E_\psi|E_\varphi)$ is $\pi' = \pi'' = 1$.

In [8, 9] it has been shown that the strict entailment satisfies the inferential rules of default logic. In general, concerning the IF-THEN rules with possibility belonging to an interval $[\pi', \pi'']$, where the extremes are lower and upper T-conditional possibilities, we could consider a set $\mathcal{D}$ including all these rules and check the inferential properties satisfied by the entailment relation as done in [3, 9].

References

[1] Bouchon-Meunier B., Coletti G., Marsala C. (2002), Independence and Possibilistic Conditioning, *Ann. of Math. and Art. Int.*, 35, pp. 107–123.

[2] Baioletti M., Coletti G., Petturiti D., Vantaggi B. (2011), Inferential models and relevant algorithms in a possibilistic framework, *Int. J. of App. Reas.*, 52(5), pp. 580–598.

[3] Coletti G., Petturiti D., Vantaggi B. (2014), Coherent T-conditional possibility envelopes and nonmonotonic reasoning, in *Inf. Proc. and Man. of Unc. in Know.-Based Sys*, Com. in Comp. and Inf. Sci., 444, (Laurent et al Eds.), Springer, pp. 446–455.

[4] Coletti G., Petturiti D., Vantaggi B. (2014), Possibilistic and probabilistic likelihood functions and their extensions: Common features and specific characteristics, *Fuzzy Sets and Sys.*, 250, pp. 25–51.

[5] Coletti G., Scozzafava R. (2004), Conditional Probability, Fuzzy Sets, and Possibility: a Unifying View, *Fuzzy Sets and Sys.*, 144, pp. 227–249.

[6] Coletti G., Scozzafava R. (2006), Conditional Probability and Fuzzy Information, *Comp. Stat. & Data Anal.*, 51, pp. 115–132.

[7] Coletti G., Scozzafava R., Vantaggi B. (2013), Inferential processes leading to possibility and necessity, *Inf. Sci.*, 245, pp. 132–145.

[8] Coletti G., Scozzafava R., Vantaggi B. (2011), Default Rules as Fuzzy Inclusion of Degree 1, *Proc. World Conf. on Soft Comp., San Francisco*, pp. 83–89.

[9] Coletti G., Scozzafava R., Vantaggi B. (2015), Possibilistic and probabilistic logic under coherence: default reasoning and System P, *Math. Slov.* (in press).

[10] Coletti G., Vantaggi B. (2009), T-conditional possibilities: Coherence and inference, *Fuzzy Sets and Sys.*, 160(3), pp. 306–324.

[11] Coletti G., Vantaggi B. (2014), Probabilistic Reasoning in a Fuzzy Context, in *Recent Dev. and New Dir. in Soft Comp.*, Stud. in Fuzz. and Soft Comp., 317, (Zadeh et al. Eds.), Sringer, pp. 97–115.

[12] de Finetti B. (1931), Sul significato soggettivo della probabilità, *Fund. Math.*, 17, pp. 298–329.

[13] Dubois D., Moral S., Prade H. (1997), A semantics for possibility theory based on likelihoods, *J. Math. Anal. Appl.*, 205, pp. 359–380.

[14] Liu Z., Li H.-X. (2005), A probabilistic fuzzy logic system for modeling and control, *IEEE Trans. on Fuzzy Sys.*, 13(6), pp. 848–859.

[15] Meghdadi A. H., Akbarzadeh-T M.-R. (2001), Probabilistic Fuzzy Logic and Probabilistic Fuzzy Systems, *Proc. IEEE Int. Fuzzy Sys. Conf.*, pp. 1127–1130.

[16] Sugeno M. (1974), *Theory of fuzzy integrals and its applications*, PhD thesis, Tokyo Institute of Technology, Tokyo, Japan.

[17] Waltman L., Kaymak U., van den Berg J. (2005), Maximum Likelihood Parameter Estimation in Probabilistic Fuzzy Classifiers, *Proc. IEEE Int. Conf. on Fuzzy Sys.*, pp. 1098–1103.

[18] Zadeh L.A. (1968), Probability measures of fuzzy events, *J. of Math. Anal. and App.*, 23, pp. 421–427.

[19] Zadeh L.A. (2002), Toward a perception-based theory of probabilistic reasoning with imprecise probabilities, *J of Stat. Plan. and Inf.*, 105, pp. 233–264.

Homomorphic Coordinates of Dempster's Semigroup

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

Abstract

Coordinates of belief functions on two-element frame of discernment are defined using homomorphisms of Dempster's semigroup (the algebra of belief functions with Dempster's rule). Three systems of the coordinates (*h-f*, *h-f*$_0$, and coordinates based on decomposition of belief functions) are analysed with a focus to their homomorphic properties. Further, ideas of generalisation of the investigated systems of coordinates to general finite frame of discernment are presented.

1 Introduction

Belief functions (BFs) are one of the widely used formalisms for uncertainty representation and processing that enable representation of incomplete and uncertain knowledge, belief updating, and combination of evidence. They were originally introduced as a principal notion of the Mathematical Theory of Evidence [14], which is often call the Dempster-Shafer Theory.

Algebraic analysis of belief functions was originally motivated by creation and analysis of combinational structure of expert systems in late 80's [10, 11]. The original algebra of belief functions with application of Dempster's rule of combination (Dempster's semigroup) was defined by Hájek and Valdés on two-element frames of discernment with elements: *Hypothesis holds, Hypothesis does not hold* [12, 13]. Some elaborations of the approach were performed by the author in early 90's.

New interest about algebraic structure related to belief functions come with investigation of conflicts of belief functions after 2010 [2, 6]. We can mention an update of older author's results on morhpisms of Dempster's semigroup [3] and first results on generalization of Dempster's semigroup to three-element frame of discernment [4, 5, 8].

At first necessary preliminaries on belief functions and basic Hájek-Valdés and author's results on Dempster's semigroup are briefly introduced (Sections 2 and 3). After that, the investigated research is presented in two parts.

The first part of this study combines both the original and new approaches. First homomorphic *h-f* coordinates come from original Hájek-Valdés results [12, 13] and

their modification h-f_0 from author's [3] (Section 4). Using [2] we can define a brand new homomorphic coordinates of Dempster's semigroup based on decomposition of BFs to their unique conflicting and non-conflicting parts (Section 5).

The second part studies issues related to generalization of the topic to three- and finite general frames of discernment (Section 6). New homomorphisms are considered. One of the results is, unfortunately, an counter-example against validity of hypothesis on unique decomposition of a BF to its conflicting and non-conflicting parts [2] in full generality. Thus the open problem of general validity of the hypothesis is transferred to the problem of finding a domain of the hypothesis validity.

2 Preliminaries

We assume classic definitions of basic notions from theory of *belief functions* [14] on finite frames of discernment $\Omega_n = \{\omega_1, \omega_2, ..., \omega_n\}$, see also [1, 8].

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 the bba are called *basic belief masses (bbm)*. $m(\emptyset) = 0$ is usually assumed. 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* $Pl(A) = \sum_{\emptyset\neq A\cap X} m(X)$. There is a unique correspondence among m and corresponding Bel and Pl thus we often speak about m as of belief function.

A *focal element* is a subset X of the frame of discernment, such that $m(X) > 0$. If all the focal elements are *singletons* (i.e. one-element subsets of Ω), then we speak about a *Bayesian belief function* (BBF); in fact, it is a probability distribution on Ω. If there are only focal elements such that $|X| = 1$ or $|X| = n$ we speak about *quasi-Bayesian BF* (qBBF). In the case of $m(\Omega) = 1$ we speak about *vacuous BF* (VBF). U_n is a BF such that $m(\{\omega_i\}) = \frac{1}{n}$ for any $1 \leq i \leq n$. A *symmetric BF* is a BF, such that $m(X) = m(Y)$ for $|X| = |Y|$, a *consonant BF* is a BF, such that its focal elements are nested, (it corresponds to necessity measure), an *exclusive BF* is a BF, such that there exists $\omega_i \in \Omega$, such that $Pl(\{\omega_i\}) = 0$, otherwise a BF is *non-exclusive*. In the case of $Pl = U_n$ we speak about a *indecicive BF*, $S_{Pl} = \{Bel \mid Pl{=}U_n\}$ is the set of all indecisive BFs.

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 [14].

Normalized plausibility of singletons[1] of Bel is a probability distribution Pl_P such that $Pl\_P(\omega_i) = \frac{Pl(\{\omega_i\})}{\sum_{\omega\in\Omega} Pl(\{\omega\})}$.

We may represent BFs by enumeration of their m-values, i.e., by (2^n–1)-tuples or by (2^n–2)-tuples as $m(\Omega_n) = 1 - \sum_{X\subsetneq\Omega_n} m(X)$; thus we have pairs (called d-pairs by Hájek & Valdés) $(a, b) = (m(\{\omega_1\}), m(\{\omega_2\}))$ for BFs on Ω_2.

[1] Plausibility of singletons is called *contour function* by Shafer in [14], thus $Pl\_P(Bel)$ is a normalization of contour function in fact.

3 Dempster's Semigroup of Belief Functions $\mathbf{D}_0$.

Hájek-Valdés algebraic structure $\mathbf{D}_0$ of non-exclusive d-pairs (i.e., exclusive pairs $(0,1)$ and $(1,0)$ are not included) with Dempster's rule $\oplus$ is called *Dempster's semigroup*, $\mathbf{D}_0 = (D_0, \oplus, -, 0, 0')$, where $0 = (0,0) = VBF$, $0' = (\frac{1}{2}, \frac{1}{2}) = U_2$, and $-(a,b) = (b,a)$, see [13]. In this study we present only several substructures related to our topic of indecisive BFs: subsemigroup of symmetric BFs $S = \{(s,s) \mid 0 \leq s \leq \frac{1}{2}\}$, and important subgroup of Bayesian BFs $G = (\{(a,b) \mid 0 \leq a,b < 1, a+b = 1\}, \oplus, -, 0')$, which is isomorphic to the additive group of reals $\mathbf{Re} = (Re, +, -, 0)$, S is isomorphic to the positive cone $\mathbf{Re}^{\geq 0}$ of $\mathbf{Re}$ extended with ∞ ($\mathbf{Re}^{+\geq 0}$). Further, we need a mapping $h(a,b) = (a,b) \oplus 0' = P\!L\!P(a,b)$ which is a homomorphic projection of the entire structure $\mathbf{D}_0$ to the group of Bayesian BFs G, i.e., $h((a,b) \oplus (c,d)) = h(a,b) \oplus h(c,d)$, where $h(a,b)$ is an abbreviation for $h((a,b))$; and a mapping $f(a,b) = (a,b) \oplus -(a,b)$ which is a homomorphic projection of $\mathbf{D}_0$ to the subsemigroup S, see Figure 1. These structures have been further studied and generalised by the author, e.g., in [1, 3, 4].

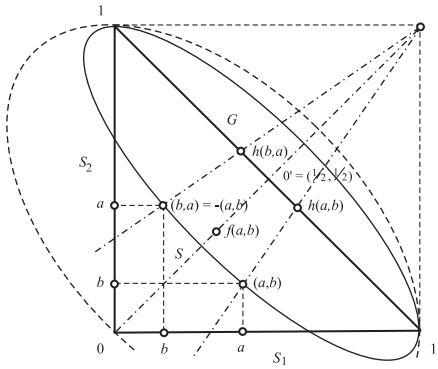


Figure 1: Dempster's semigroup $\mathbf{D}_0$. Homomorphism h is in this representation a projection of the triangle representing D_0 to its hypotenuse G along the straight lines running through the point $(1,1)$. All of the d-pairs lying on the same ellipse (running through points $(0,1)$ and $(1,0)$) are mapped by f to the same $f(a,b) \in S$.

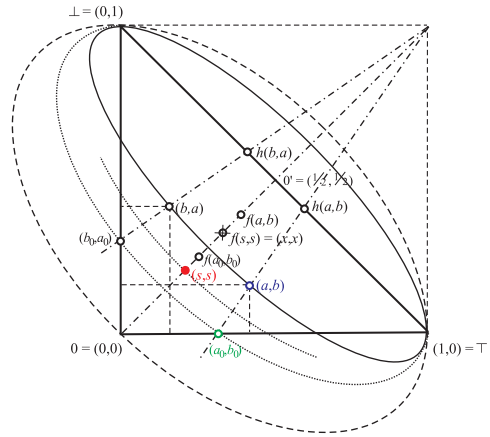


Figure 2: Non-conflicting part (a_0, b_0) and conflicting part (s,s) of a BF (a,b) on a 2-element frame of discernment Ω_2: $(a,b) = (a_0,b_0) \oplus (s,s)$.

Theorem 1 *Any BF (a,b) on a 2-element frame of discernment Ω_2 is Dempster's sum of its unique* non-conflicting part $(a_0,b_0) \in S_1 \cup S_2$ *and of its unique* conflicting part $(s,s) \in S$, *which does not prefer any element of Ω_2, i.e., $(a,b) = (a_0,b_0) \oplus (s,s)$, see Figure 2. It holds true that $s = \frac{b(1-a)}{1-2a+b-ab+a^2} = \frac{b(1-b)}{1-a+ab-b^2}$ and $(a,b) = (\frac{a-b}{1-b}, 0) \oplus (s,s)$ for $a \geq b$; and similarly that $s = \frac{a(1-b)}{1+a-2b-ab+b^2} = \frac{a(1-a)}{1-b+ab-a^2}$ and $(a,b) = (0, \frac{b-a}{1-a}) \oplus (s,s)$ for $a \leq b$.*

4 *h-f* Homomorphic Coordinates of $\mathbf{D}_0$

Taking homomorphisms h and f we can define bijection λ of $\mathbf{D}_0$ into $G \times S$: $\lambda(a,b) = (h(a,b), f(a,b))$; from [12, 13] we have the following theorem:

Theorem 2 (Hájek-Valdés) *(i) The mapping λ associating with each d-pair (a,b) the pair $\lambda(a,b) = (h(a,b), f(a,b))$ is one-one mapping of $\mathbf{D}_0$ into $G \times S$.*
(ii) λ is not onto: in fact, given $(g, 1-g) \in G$ and $(s,s) \in S$ there is a $(a,b) \in \mathbf{D}_0$ such that $h(a,b) = (g, 1-g)$ and $f(a,b) = (s,s)$ iff

$$(1-2s)/(2-3s) \leq g \leq (1-s)/(2-3s). \tag{1}$$

Each d-pair from $\mathbf{D}_0$ is uniquely defined by a corresponding pair $\lambda(a,b) = (h(a,b), f(a,b))$, thus $h(a,b)$ and $f(a,b))$ are coordinates of $\mathbf{D}_0$. Both of the singular coordinates are homomorphisms of $\mathbf{D}_0$ onto G (or S, respectively), thus we speak about *homomorphic coordinates*. Moreover λ is homomorphisms of $\mathbf{D}_0$ into $(G \times S, \boxplus, \boxminus, (0',0), (0',0'))$, where operations $\boxplus$ and $\boxminus$ are given by $\oplus$ and $-$ application coordinate by coordinate on $G \times S$, see Thm. 6.39 in [15].

Similarly to homomorphism f there is homomorphism f_0 of $\mathbf{D}_0$ onto S [3] such that $f_0(a,b) \oplus f_0(a,b) = f(a,b)$, thus $f_0(a,b)$ is Dempster's 'half' of $f(a,b)$.

All d-pairs from an ellipse going through points $(0,1)$ and $(1,0)$ have same $f(a,b) = (s,s)$. Formally the corresponding part of the ellipse is given by $\{(d_1,d_2) \mid (d_1,d_2) \in D_0, sd_1^2 + sd_2^2 + d_1d_2 - d_1 - d_2 + s = 0\}$, see Lemma 6.33 in [15]. All these d-pairs have also same $f_0(a,b) = (s_0,s_0)$ which is intersection of the ellipse with S ($(s_0,s_0) \oplus (s_0,s_0) = (s,s)$). Thus homomorphism f_0 has more intuitive interpretation than homomorphism f has.

Analogously to definition of λ we can define bijection λ_0 of $\mathbf{D}_0$ into $G \times S$ using homomorphisms h and f_0: $\lambda_0(a,b) = (h(a,b), f_0(a,b))$. We can easily verify that λ_0 has same algebraic properties as λ has. Thus we obtain:

Theorem 3 *(i) The mapping λ_0 associating with each d-pair (a,b) the pair $\lambda_0(a,b) = (h(a,b), f_0(a,b))$ is one-one mapping of $\mathbf{D}_0$ into $G \times S$.*
(ii) λ_0 is not onto: in fact, given $(g, 1-g) \in G$ and $(s,s) \in S$ there is a $(a,b) \in \mathbf{D}_0$ such that $h(a,b) = (g, 1-g)$, $f_0(a,b) = (s_0,s_0)$, and $f(a,b) = (s,s)$

iff
$$\frac{1-2s}{2-3s} \leq g \leq \frac{1-s}{2-3s},$$

iff
$$\frac{1-4s_0+4s_0^2}{2-6s_0+5s_0^2} \leq g \leq \frac{1-2s_0+s_0^2}{2-6s_0+5s_0^2}.$$

Proof. (i) and the first equivalence of (ii) is just verification that f_0 has the same algebraic properties as f has. The second equivalence comes from substitution $s = \frac{2s_0-3s_0^2}{1-2s0^2}$, where $(s_0,s_0) \oplus (s_0,s_0) = (\frac{2s_0-3s_0^2}{1-2s0^2}, \frac{2s_0-3s_0^2}{1-2s0^2}) = (s,s)$. □

We have isomorphisms G onto $(Re, +, -, 0)$ and S onto $(Re^{+\geq 0}, +, 0, \infty)$ [12, 13, 15]. Hence isomorphism of $\mathbf{D}_0$ into $(Re, +, -, 0) \times (Re^{+\geq 0}, +, 0, \infty)$ follows isomorphisms λ, λ_0. From this we can easily see strength of Bayesisan BFs (absorbing

property), due to they have f (and f_0) coordinates equal to ∞. Analogously we can see that a 'small' difference from Bayesian BFs and from exclusive (extremal) BFs $(0,1)$ and $(1,0)$ is greater then relatively same difference deep inside the D_0 triangle or close to 0, see also the following simple example.

Example 1. Let suppose the following simple d-pairs:
$(a_{11},0)=(0.1,0)$, $(a_{21},0)=(0.97,0)$, $(a_{31},0)=(0.99999,0)$,
$(a_{12},0)=(0.2,0)$, $(a_{22},0)=(0.98,0)$, $(a_{32},0)=(0.999999,0)$.
We obtain the following h projections:
$h(a_{11},0)=(0.5263,0.4737)$, $h(a_{12},0)=(0.5555,0.4444)$, which are mapped to $(1.668, \infty)$ and $(1.807,\infty)$, thus difference 0.1 of the first component is mapped by isomorphism of $G^{\geq 0'}$ onto $Re^{\geq 0}$ $ln(\frac{1+x}{1-x})$, see [9], to similar difference 0.139;
$h(a_{21},0)=(0.970874,0.029126)$, $h(a_{22},0)=(0.980392,0.019608)$, which are mapped to $(6.080,\infty)$ and $(6.658,\infty)$, thus difference 0.01 of the first component is mapped to greater difference 0.658;
$h(a_{31},0)=(0.99999,0)$, $h(a_{32},0)=(0.999999,0)$, which are mapped to $(17.609,\infty)$ and $(20.931,\infty)$, thus difference 0.000009 of the first component is mapped to significantly greater difference 3.322.

From the example we see, that we have to be careful when using values close to 1 and analogously, such that sum of all values is close to 1 (i.e., BFs is close to Bayesian BFs) and that we have to use enough precise computing, as e.g., the at computation using only 6 decimal digits, 22 is not distinguishable from ∞ when d-pairs are mapped to $Re \times Re^{+\geq 0}$ (due to $0.9999995 \sim 21.93157$).

This is also related to *realizations* of user pairs (given by a finite user scale $-N,-N+1,...,-1,0,1,2,...N$) in D_0, see [10, 11, 12, 13] from the period when WUPES workshop was established.

5 Coordinates of $\mathbf{D_0}$ Based on Conflicting and Non-Conflicting Parts

We have unique decomposition of a BF on a two-element frame of discernment to its conflicting and non-conflicting parts: $Bel = Bel_0 \oplus Bel_S$). This decomposition was derived using homomorphic properties of mappings h and f. Moreover any pair of simple Bel_0 and symmetric Bel_S defines a BF on Ω_2 and any BFs on Ω_2 is decomposable. Thus we have a bijection between D_0 and $S_1 \cup S_2 \times S$, using a one-one correspondence between S_1 and $G^{\geq 0'}$ and between S_2 and $G^{\leq 0'}$, hence we obtain a bijection between $S_1 \cup S_2$ and G and the following lemma:

Lemma 1 *The mapping κ associating with each d-pair $Bel=(a,b)$ the pair $\kappa(Bel) = (h(Bel_0), Bel_S) = (h(Bel), Bel_S)$, where $Bel_0 \oplus Bel_S = Bel$ is decomposition of Bel into its conflicting and non-conflicting parts, is one-one mapping of $\mathbf{D}_0$ onto $G \times S$.*

The bijection κ is constructed using homomorphic properties of mappings h and f, is it also homomorphisms itself? Does it hold true that, for $Bel' = Bel'_0 \oplus Bel'_S$

and $Bel' = Bel'_0 \oplus Bel'_S$ there is $Bel = Bel' \oplus Bel'' = Bel_0 \oplus Bel_S$ such that $Bel_0 = Bel'_0 \oplus Bel''_0$ (thus also $h(Bel) = h(Bel') \oplus h(Bel'')$) and $Bel_S = Bel'_S \oplus Bel''_S$? Unfortunately, we can easily find a counter-example.

Example 2. Let us suppose for simplicity two simple d-pairs $Bel' = (0.6, 0) \in S_1$ and $Bel'' = (0, 0.2)$, thus we have trivial decompositions, where $Bel^i_0 = Bel^i$ and $Bel^i_S = (0, 0)$. We have $Bel = Bel' \oplus Bel'' = (0.6, 0) \oplus (0, 0.2) = (\frac{6}{11}, \frac{1}{11})$ which does not correspond to non-conflicting part, thus we obtain: $Pl = (\frac{10}{11}, \frac{5}{11})$, $Bel_0 = (\frac{5}{10}, 0)$ and $Bel_S = (\frac{\frac{1}{11} \cdot \frac{5}{11}}{1 - 2\frac{6}{11} + \frac{1}{11} - \frac{6}{11} \cdot \frac{1}{11} + \frac{6}{11}\frac{6}{11}}, \frac{5}{121 - 22 \cdot 6 + 11 - 6 + 36}) = (\frac{1}{6}, \frac{1}{6}) \neq (0, 0)$.

This comes from the fact that Bel' and Bel'' are conflicting thus internal conflict is increased; We have $Bel' = (0.6, 0) = (0.5, 0) \oplus (0.2, 0)$ thus $Bel' \oplus Bel'' = (0.5, 0) \oplus (0.2, 0) \oplus (0, 0.2)$, where $(0.5, 0) = Bel_0$ and $(0.2, 0) \oplus (0, 0.2) = (\frac{16}{96}, \frac{16}{96}) = (\frac{1}{6}, \frac{1}{6})$ produces an increase of internal conflict: $Bel_S = Bel'_S \oplus Bel''_S \oplus (0.2, 0) \oplus (0, 0.2) = (0, 0) \oplus (0, 0) \oplus (\frac{1}{6}, \frac{1}{6}) = (\frac{1}{6}, \frac{1}{6})$.

We obtain analogous results whenever one of d-pairs is in S_1 and the other in S_2 and more generally when one of d-pairs is in $D_0^{\geq 0} \setminus S$ and the other in $D_0^{\leq 0'} \setminus S$. What about a situation, when both d-pairs are in the same half of the D_0 triangle, e.g., $Bel', Bel'' \in D_0^{\geq 0}$?

Let us have $Bel' = (a, b)$, $Bel'' = (c, d)$, where $a \geq b$ and $c \geq d$. Thus $Bel' = Bel'_0 \oplus Bel'_S = (\frac{a'-b'}{1-b'}, 0) \oplus (s', s')$ and analogously $Bel' = (\frac{a''-b''}{1-b''}, 0) \oplus (s'', s'')$. Using this we obtain $Bel = Bel' \oplus Bel'' = (\frac{a'-b'}{1-b'}, 0) \oplus (s', s') \oplus (\frac{a''-b''}{1-b''}, 0) \oplus (s'', s'') = (\frac{a'-b'}{1-b'} + \frac{a''-b''}{1-b''} \cdot \frac{1-a'}{1-b'}, 0) \oplus (s', s') \oplus (s'', s'')$. Hence $Bel_0 = Bel'_0 \oplus Bel''_0 = (\frac{a'-b'}{1-b'} + \frac{a''-b''}{1-b''} \cdot \frac{1-a'}{1-b'}, 0)$ and $Bel_S = Bel'_S \oplus Bel''_S = \oplus(s', s') \oplus (s'', s'')$. The situation when both d-pairs are in the left cone of D_0 triangle is analogous. Thus the decomposition of a d-pair to its conflicting and non-conflicting parts does not commute with Dempster's rule $\oplus$ on the entire D_0, but it commutes with $\oplus$ on both the cones of D_0: $D_0^{\geq 0}$ and $D_0^{\leq 0'}$. Thus we have homomorphicity of the coordinates based on conflicting and non-conflicting parts separately on both the cones of D_0:

Theorem 4 *(i) The mapping κ associating with each d-pair $Bel = (a, b)$ the pair $\kappa(Bel) = (h(Bel_0), Bel_S) = (h(Bel), Bel_S)$, where $Bel_0 \oplus Bel_S = Bel$ is decomposition of Bel into its conflicting and non-conflicting parts, is one-one mapping of $\mathbf{D}_0$ onto $G \times S$.*
(ii) The mapping κ does not commute with Dempster's rule $\oplus$ on entire $\mathbf{D}_0$, but it does on both its cones $\mathbf{D}_0^{\geq 0}$ and $\mathbf{D}_0^{\leq 0'}$.

Using the homomorphic coordinates we can generalise the operation of *Dempster's half* from G and S onto entire $\mathbf{D}_0$. We derived empirically Dempster's half for d-pairs from S in [2], in fact it corresponds to half in reals as S is isomorphic to the positive cone of the extended additive group of reals $\mathbf{R}e^{+\geq 0} = (Re^{\geq 0} \cup \{\infty\}, +, 0)$, in the same way it holds for group G, which is isomorphic to the entire additive group of reals $\mathbf{R}e$.

For general d-pair (a, b) and $(a, b) \oplus (a, b)$ are always in the same cone, thus also any d-pair and its Dempster's half are in the same cone of $\mathbf{D}_0$, thus we can use

the homomorphic properties of mapping κ. Thus for any d-pair $Bel = (a, b)$ we have $h(a, b) \in G$ and $Bel_S \in S$, thus we have Dempster's halves ${}'\frac{1}{2}'(h(a, b)) \oplus' \frac{1}{2}'(h(a, b)) = h(a, b)$, ${}'\frac{1}{2}'(Bel_S) \oplus' \frac{1}{2}'(Bel_S) = Bel_S$. Using them we obtain ${}'\frac{1}{2}'(a, b) = \kappa^{-1}({}'\frac{1}{2}'(h(a, b)),' \frac{1}{2}'(f(a, b)))$. Geometrically we project (a, b) to G and S (to $h(a, b)$ and $(a, b)_S$), make Dempster's halves there and the result is just the intersection of preimages of separate halves, which is unique in D_0 using bijectivity of κ. Thus we have proved the following lemma:

Lemma 2 *For any belief function Bel on a two-element frame there exists its Dempster's half, i.e., belief function Bel′ such that Bel′ ⊕ Bel″ = Bel.*

Alternatively, we can use for a proof also $h - f$ coordinates and verify condition for preimage. The idea is simple but the formulas are rather complicated.

The relatively simple proof of Dempster's half existence using coordinates based on conflicting parts motivates us to show analogously also subtraction on $\mathbf{D}_0$, but it is not possible. Let suppose again BFs Bel', Bel'' from $\mathbf{D}_0 \geq 0$: having $h(Bel') \leq h(Bel'')$ and $Bel'_S \geq Bel''_S$, we have $Bel'_S = Bel''_S \oplus (x, x)$, where $(x, x) \in S$ and $h(Bel') = h(Bel'') \oplus (y, 1 - y)$, but $(y, 1 - y) \in G^{\leq 0'}$ due to $h(Bel') \leq h(Bel'')$, thus $h(Bel') \in \mathbf{D}_0^{\leq 0'}$, hence we cannot use homomorphic property of κ which holds on both the cones of $\mathbf{D}_0$ separately. (We need to use strict inequalities for creating of an counter-example).

We have not such a problem with subtractions on G using $h - f$ coordinates, but there can arise a problem with non-existence of λ preimage. Let us suppose Bel', Bel'' from $\mathbf{D}_0^{\geq 0}$ such $h(Bel') \leq h(Bel'')$ and $f(Bel') = f(Bel')$ now. In this case $Bel' \oplus (x, y) = Bel''$, where $f(x, y) = (0, 0)$, but there is the only λ-preimage $(0, 0)$ for coordinates $((a, b), (0.0))$: for $\lambda^{-1}((a, b), (0.0)) = (0, 0)$, thus it must be $h(Bel') = h(Bel'')$, thus also $Bel' = Bel''$. Thus both $Bel' \oplus Bel_x = Bel''$ and $Bel'' \oplus Bel_y = Bel'$ have solution on $\mathbf{D}_0$ iff $Bel' = Bel''$.

Fact 1 *There is no subtraction for a general couple of BFs either on* $\mathbf{D}_0$ *or on its cones* $\mathbf{D}_0^{\geq 0}$, $\mathbf{D}_0^{\leq 0'}$.

6 Towards Homomorphic Coordinates on a General Finite Frame of Discernment

A general case of belief functions on a general finite frame $\Omega_n = \{\omega_1, \omega_2, ..., \omega_n\}$ is more complex. There is both qualitative and quantitative increase of complexity. BFs on a two-element frame Ω_2 are simply representable by pairs from 2-dimensional structure. All BFs on Ω_2 are quasi-Bayesian: there are only singleton focal elements and the entire frame. Quasi-Bayesian BFs on Ω_n are representable by n-tuples from n-dimensional structure, there is only linear quantitative increase of complexity. For representation of general BFs on Ω_n we need $2^n - 2$-tuples from a $2^n - 2$-dimensional structure (remaining m-value $m(\Omega_n)$ may be computed as $1 - \sum_{X \subsetneq \Omega} m(X)$), thus

there is exponential increase of quantitative complexity and there is also structural increase of complexity as the dimension corresponding to a proper subset A of Ω_n is somehow related to the dimensions related to all elements of A, to dimensions related to all subsets of A, to dimensions related to all supersets of A, and generally somehow related to all dimensions related all $B \subset \Omega_n$ such that $A \cap B \neq \emptyset$. The dimensions are not related, they are orthogonal for any couple of disjoint sets $B \cap C = \emptyset$.

Due to the complexity first attempts to generalisation of algebraic structures have been performed on a three-element frame of discernment Ω_3. Bayesian and indecisive BFs on Ω_n have been recently presented in [7].

Let us denote by D_n^+ the set of all $2^n - 2$-tuples representing BFs on Ω_n thus $D_n^+ = \{(d_1, d_2, ..., d_n, d_{12}, ..., d_{234...n}) \mid \sum_{X \subset \Omega_n} d_X \leq 1\}$, where $d_X = m(X)$ such that X in index of d_X is the list of indices of ω_i's contained in $X \subset \Omega_n$. Further $D_n = \{(d_1, d_2, ..., d_{23...n}) \mid Pl(\omega_i) > 0 \text{ for all } 1 \leq i \leq n\}$ is the set of all non-exclusive BFs ($2^n - 2$-tuples), thus Dempster's combination $\oplus$ is defined for any couple of tuples from D_n. Then $\mathbf{D}_n = (D_n, \oplus, 0, U_n)$ is a partial generalisation of $\mathbf{D}_0$ as operation '$-$' is still not defined.

All three systems of coordinates presented in the previous sections $\lambda(a,b) = (h(a,b), f(a,b))$, $\lambda_0(a,b) = (h(a,b), f_0(a,b))$, and $\kappa(a,b) = (h(a,b), (a,b)_S)$ have the $h(a,b)$ as their first component. There is the generalisation of h to BFs on Ω_n, $h : D_n \longrightarrow G_n$, $h(Bel) = Bel \oplus U_n = Pl\!\!\perp\!\!P$, see [2]. It is a homomorphic projection of $\mathbf{D}_n$ to G_n, where G_n is group of non-excluding Bayesian BFs [7].

Significantly more complicated is a situation with the second coordinates. We have not yet a generalisation of the operation '$-$' thus either of homomorphisms f and f_0. There is only generalisation of '$-$' on Bayesian BFs and on some other special cases. There is a homomorphism f_π which maps $\mathbf{D}_n$ to the structure of all symmetric BFs, but this seems not to be a actual generalisation of f, see [6]. This topic is still under development.

There is the following hypothesis on decomposition of a general BF, see [2]; existence of the non-conflicting part was proven in general, existence of conflicting part only for BFs on Ω_2.

Hypothesis 1 *We can represent any BF Bel on n-element frame of discernment Ω_n as Dempster's sum $Bel = Bel_0 \oplus Bel_S$ of non-conflicting consonant BF Bel_0 and of indecisive conflicting BF Bel_S which has no decisional support, i.e. which does not prefer any element of Ω_n to the others, i.e., $Pl\!\!\perp\!\!P(\omega_i) = \frac{1}{n}$.*

Assuming this hypothesis we can make some investigation of coordinates based on decomposition of BFs. Having $Bel = Bel_0 \oplus Bel_S$, we have again $\kappa(Bel) = (h(Bel), Bel_S)$, where Bayesian $h(Bel)$ uniquely corresponds to consonant Bel_0. We can show that $\kappa(Bel)$ really determines $2^n - 2$ coordinates in simplex of BFs on Ω_n. In detail, $h(Bel)$ determines $n-1$ coordinates and remaining $2^n - (n+1)$ coordinates are determined by Bel_S:

We have up to n singleton focal elements at $h(Bel)$ in general, resp. just n singleton focal elements for non-exclusive BFs. This corresponds to $n-1$ dimensional structure, because one of the dimensions is redundant: $m(\{\omega_n\}) = 1 - \sum_1^{n-1} m(\{\omega_i\})$ for Bayesian BFs. In figures of $\mathbf{D}_3$ this corresponds to the 2-dimensional triangle

$((1,0,0,0,0,0),(0,1,0,0,0,0),(0,0,1,0,0,0))$ for BFs on Ω_3. (this should hold also for potential generalisation of f-h coordinates).

$Bel_S \in S_{Pl} = \{Bel \mid Pl = U_n\}$ have 2^n-1 positive focal elements, but it is only $2^n-(n+1)$ dimentisons as $m(\emptyset)=0$ and $m(\Omega_n)$ is $1-\sum_{X\subsetneq\Omega} m(X)$ as usually. And further $n-1$ focal elements are determined by the others as it should hold that $h(Bel_S)=U_n$, thus all singletons should have the same plausibility. We can freely set m-values for non-singletons (proper subsets of Ω_n) and commute $Pl_w(\{\omega_i\}) = \sum_{\omega_i\in X, 1<|X|<n} m(X)$ after we can freely set value of one of singletons, let us say ω_1, and compute $m(\{\omega_j\}) = Pl_w(\{\omega_1\}) + m(\{\omega_1\}) - Pl_w(\{\omega_j\})$ and finally compute $m(\Omega_n)$.[2]

Example 3. Let us assume a BF Bel on Ω_3 given by 6-tuple if its m values $(\frac{19}{38}, \frac{7}{38}, \frac{4}{38}, \frac{2}{38}, 0, \frac{2}{38}; \frac{4}{38})$ where $\frac{4}{38}$ is the 7th redundant value of $m(\Omega_3)$. We have $h(Bel) = (0.5, 0.3, 0.2, 0, 0, 0; 0)$ and decomposition $Bel = Bel_0 \oplus Bel_S = (0.4,0,0,0.2,0,0;0.4) \oplus (0.3,0.2,0.2,0,0,0.1;0.2)$. Thus we have 2 dimensions given by $h(Bel)(\{\omega_1\}) = 0.5$ and $h(Bel)(\{\omega_2\}) = 0.3$ (the third value of $h(Bel)$ is redundant $1-0.5-0.3 = 0.2$) and 4 dimensions given by $Bel_S = (0.3, 0.2, 0.2, 0, 0, 0.1; 0.2)$: $m_S(\{\omega_1,\omega_2\}) = m_S(\{\omega_1,\omega_3\}) = 0$, $m_S(\{\omega_2,\omega_3\}) = 0.1$, and $m_S(\{\omega_1\}) = 0.3$, 3 other values are redundant, thus computed:
$m_S(\{\omega_2\}) = (m_S(\{\omega_1,\omega_2\}) + m_S(\{\omega_1,\omega_3\}) + m_S(\{\omega_1\}) - (m_S(\{\omega_1,\omega_2\}) + m_S(\{\omega_2,\omega_3\}) = (0+0+0.3) - (0+0.1) = 0.3 - 0.1 = 0.2$, $m_S(\{\omega_3\}) = 0.3 - 0.1 = 0.2$ and $m(\Omega_n) = 1 - (0.3+0.2+0.2+0+0+0.1) = 0.2$.

6.1 Homomorphicity of coordinates based on decomposition of BFs

We have seen that bijection κ is not homomorphism of entire $\mathbf{D}_0$ but of its two cones $\mathbf{D}_0^{\geq 0}$, $\mathbf{D}_0^{\leq 0'}$, it is caused by appearing of internal conflict of result when two mutually conflicting consonant non-conflicting parts of conflicting BFs are combined. Analogously conflict appears when two non-conflicting parts with their disjoint least focal elements are combined. What about a situation where two BFs have non-conflicting parts with same least focal elements? There is no conflict between them. Thus there is no conflict among non-conflicting parts all BFs which have same least focal element of their non-conflicting parts, thus the same element with max contour (plausibility of singleton). We have n such sets for n singletons. Is mapping κ homomorphism of such sets? On three element frame such sets correspond to subsimplices S_{mi} where, e.g., $S_{m1} = (\{(d_1,d_2,d_3,d_{12},d_{13},d_{23}) \in D_3 \mid Pl(\{\omega_1\}) \geq Pl(\{\omega_2\}), Pl(\{\omega_3\})\}, \oplus, 0\}$, we can show that is $(S_{m1}, \oplus, 0, U_3)$ is subalgebra of $\mathbf{D}_3$. Thus algebras of BFs with maximal contour reached by ω_i, let us note that these subalgebras of $\mathbf{D}_0$ are out of scope of [5, 8]. Unfortunately, max contour for the same element is not enough for homomorphicity of κ:

[2] This is just a sketch of an algorithm for generating of a BF $Bel \in S_{Pl}$. For a complete algorithm we should keep some inequalities to obtain all m-values between 0 and 1 summing up to 1 or select the singleton with maximal Pl_w value instead of ω_1 for free setting of m and finally make normalisation of the values.

Example 4. Let us suppose any BFs Bel' and Bel'' such that $Bel'_0 = (0.1, 0, 0, 0.8, 0, 0; 0.1)$ and $Bel''_0 = (0.1, 0, 0, 0, 0.6, 0; 0.3)$, their non-conflicting parts are mutually non-conflicting as intersection of all their focal elements is non-empty (equal to $\{\omega_1\}$), $Bel'_0 \oplus Bel''_0 = (0.67, 0, 0, 0.24, 0, 06, 0; 0.03)$ (as all focal elements are mutually non-disjoint no normalisation is used), unfortunately the result is not consonant, thus it is not equal to $(Bel' \oplus Bel'')_0 = (Bel'_0 \oplus Bel''_0)_0 = (0.73, 0, 0, 0.18, 0, 0; 0.09)$. Hence also $Bel'_S \oplus Bel''_S \neq (Bel' \oplus Bel'')_S$ as some internal conflict has arisen at combining $Bel'_0 \oplus Bel''_0$ thus conflicting part has increased.

Using the example we can see that we need some stronger condition for homomorphicity of $\kappa(Bel) = (h(Bel), Bel_S)$ resp. of mapping f_S such that $f_S(Bel) = Bel_S$. Let us suppose BFs such that orderings of ω_i's according their plausibility (resp. contour) values are the same. We can denote it S_{o_X}, where X is an code for ordering of elements according their contour value. E.g., $S_{o_{123}}$ is set of all BFs on Ω_3 with max contour at ω_1 and minimal at ω_3, $S_{o_{213}}$ is set of all BFs on Ω_3 with max contour at ω_2 and minimal at ω_3, etc. It is possible prove that the $(S_{o_X}, \oplus, 0, U_n)$ is an subalgebra of $\mathbf{D}_n$ if non-strict ordering is considered, because 0 and U_3 and all other symmetric BFs are in any S_{o_X} as any of their elements have max contour and min contour in the same time. There are $n!$ such subalgebras S_{o_X}, as there are $n!$ orderings of n elements: n elements may be the first, $n-1$ elements may be the second, when the first is already fixed, $n-2$ elements may be the third, when the first two are already fixed, ..., 2 elements may be the last but one, when order of $n-2$ elements is already fixed, and the only element may be the last, when order of all others is already fixed. Thus for Ω_2 we have $2! = 2$ orderings of elements, which correspond to two cones $\mathbf{D}_0^{\geq 0}$, $\mathbf{D}_0^{\leq 0'}$. We have $3! = 6$ subalgebras of $\mathbf{D}_3$ of BFs with same ordering of their contour values.

Lemma 3 *(i) The mapping f_S assigning to a BF Bel its conflicting part $f_S(Bel) = Bel_S$ commutes with $\oplus$ on subalgebras S_{o_X} of BFs with fixed order of elements according to their contour values. I.e., f_S is homomorphism of intersection of its definition domain with any S_{o_X}, thus of the structures $(Dom(f_S) \cap S_{o_X}, \oplus, 0, U_n)$.*
(ii) For any BF from definition domain $(Dom(f_S)$ of f_S there exits its Dempster's half $Bel_{1/2}$, such that $Bel_{1/2} \oplus Bel_{1/2} = Bel$, whenever there exists Dempster's half on algebra of indecisive BFs S_{Pl}.

Proof. (i) Non-conflicting parts of BFs with the same order of contour values have the same focal elements, thus $Bel'_0 \oplus Bel''_0$ has the same focal elements as Bel'_0 and $Bel"_0$ (as their focal elements are nested), thus it is consonant hence equal to $(Bel'_0 \oplus Bel''_0)_0 = (Bel' \oplus Bel'')_0$. Hence also $Bel'_S \oplus Bel''_S = (Bel' \oplus Bel'')_S$. If order of contour values is not strict, then sets of focal elements of Bel'_0, Bel''_0, and $Bel'_0 \oplus Bel"_0$ are subset of the set from the strict case.
(ii) Analogously to two-element frame, any BF and its Dempster's half are in the same homomorphic subalgebra, thus we can use homomorphic property of f_S. $Bel_{1/2}$ is an intersection of preimages of Dempster's halves of Bel_S (if it exists) and of $h(Bel)$. (Dempsters's half on G_n: $(\frac{\sqrt{d_1}}{\sqrt{d_1}+\sqrt{d_2}+...+\sqrt{d_n}}, \frac{\sqrt{d_2}}{\sum_1^n \sqrt{d_i}}, ..., \frac{\sqrt{d_n}}{\sum_1^n \sqrt{d_i}}, 0, 0, ..., 0) \oplus (\frac{\sqrt{d_1}}{\sum_1^n \sqrt{d_i}}, \frac{\sqrt{d_2}}{\sum_1^n \sqrt{d_i}}, ..., \frac{\sqrt{d_n}}{\sum_1^n \sqrt{d_i}}, 0, 0, ..., 0) = (d_1, d_2, ..., d_n, 0, 0, ..., 0))$. □

Example 4. (cont.) We have $Bel'_0 \oplus Bel''_0$ and $(Bel'_0 \oplus Bel''_0)_0$, thus it should be $(0.73,0,0,0.18,0,0;0.09) \oplus (x_1,x_2,x_3,x_{12},x_{13},x_{23};x) = (0.67,0,0,0.24,0.06,0;0.03)$; from this we obtain equations: $\frac{73}{100}(x_1+x_{12}+x_{13}+x)+\frac{18}{100}(x_1+x_{13})+\frac{9}{100}x_1 = \frac{67}{100k}$, $\frac{18}{100}(x_2+x_{23})+\frac{9}{100}x_2 = 0$, $\frac{9}{100}x_2 = 0$, $\frac{18}{100}(x_{12}+x)+\frac{9}{100}x_{12} = \frac{24}{100k}$, $\frac{9}{100}x_{13} = \frac{6}{100k}$, $\frac{9}{100}x_{23} = 0$, $\frac{9}{100}x = \frac{3}{100k}$. Solving these equations we obtain $k = 1$ and $(x_1,x_2,x_3,x_{12},x_{13},x_{23};x) = (-\frac{2}{3},0,0,\frac{2}{3},\frac{2}{3},0;\frac{1}{3})$.
We have confirmed existence of $(Bel'_0 \oplus Bel''_0)_S$, on the other hand we see that it is not a BF according the classic Shafer's definition, but some generalised one.

Example 5. Let us consider BFs Bel', Bel''' with same order of contour value such that $Bel'_0 = (0.1,0,0,0.8,0,0;0.1)$ and $Bel'''_0 = (0.1,0,0,0.6,0,0;0.3)$. $Bel'_0 \oplus Bel'''_0 = (0.19,0,0,0.78,0,0;0.03)$, it is consonant thus $Bel'_0 \oplus Bel'''_0 = (Bel' \oplus Bel''')_0$. From $Bel' \oplus Bel''' = Bel'_0 \oplus Bel'''_0 \oplus Bel'_S \oplus Bel'''_S$ we obtain also $Bel'_S \oplus Bel'''_S = (Bel' \oplus Bel''')_S$.

6.2 New open problems

From the previous example we can see, that the Hypothesis 1 does not hold true in general. Thus instead of question of validity of the hypothesis, we obtain new open problems: For which set of BFs Hypothesis 1 holds true? What are the generalised belief functions, which are conflicting parts of BFs out of validity of Hypothesis 1. For which set of generalised BFs Hypothesis 1 holds true?

7 Conclusion

Three systems of coordinates of belief functions on a two-element frame of discernment were defined and analysed: *h-f* coordinates are defined using homomorphic projection of Dempster's semigroup, unfortunately, there are also coordinates which do not correspond to any belief functions; *h-f*$_0$ coordinates have more intuitive interpretation and the same algebraic properties as *h-f* coordinates have, coordinates based on the decomposition of belief functions into their unique conflicting and non-conflicting parts are homomorphic on two cones of Dempster's semigroup separately, on the other hand there exists a belief function to any coordinates and vice-versa.

Ideas of generalisation of the presented systems of coordinates to belief functions on general finite frame of discernment have been analysed and several partial results relating to generalisations have been presented. Among these results is an example of a belief function which has not its decomposition into conflicting and non-conflicting parts in the domain of classical belief functions. Thus open problem of existence of the decomposition was change to problem of domain of decomposition, and problem of existence of decomposition into generalised belief functions and a way how belief functions have to be generalised.

References

[1] Daniel M. (1995), Algebraic structures related to Dempster-Shafer theory. In: Bouchon-Meunier B. et al. (eds.) *Advances in Intelligent Computing - IPMU'94*. LNCS vol. 945, Springer, Berlin Heidelberg, 51–61.

[2] Daniel M. (2011), Non-conflicting and Conflicting Parts of Belief Functions. In: Coolen, F., de Cooman, G., Fetz, T., Oberguggenberger, M. (eds.) *ISIPTA'11; Proceedings of the 7th ISIPTA*. SIPTA, 139–148.

[3] Daniel M. (2011), Morphisms of Dempster's Semigroup: A Revision and Interpretation. In: Barták, R. (ed.) *Proc. of 14th Czech-Japan Seminar on Data Analysis and ... Uncertainty, CJS 2011*. Matfyzpress, Prague, 26–34.

[4] Daniel M. (2012), Introduction to an Algebra of Belief Functions on Three-element Frame of Discernment — a Quasi Bayesian Case. In: S. Greco et al. (eds.) *IPMU 2012, Part III.* CCIS, vol. 299, Springer, Heidelberg, 532–542.

[5] Daniel, M. (2012), Introduction to an Algebra of Belief Functions on Three-element Frame of Discernment — a General Case. In: Kroupa, T., Vejnarová, J. (eds.), Proc. WUPES 2012, Univ. of Ecomomics Prague, 46–57.

[6] Daniel M. (2014), Towards a Conflicting Part of a Belief Function. In: Laurent A et al. (eds.), *Proc. of IPMU 2014*. Communications in Computer and Information Science, vol. 444, Springer, Berlin Heidelberg, 212–222.

[7] Daniel M. (2015), Indecisive Belief Functions. In: *Proc. of 18th Czech-Japan Seminar on Data Analysis and ... Uncertainty, CJS 2015*. In print.

[8] Daniel M. (2015), Basic Algebraic Structures Related to Belief Functions on a Three-element Frame of Discernment. (Subm. to Fuzzy Sets and Sys).

[9] Hájek,P. (1985), Combining functions for certainty degrees in consulting systems. Int. J. Man-Machine Studies, **22** (1): 59–76.

[10] Hájek,P., Hájková,M. (1990), The Expert System Shell EQUANT-PC: Philosophy, Structure and Implementation. In: Computational Statistics Quarterly, **4**: 261–267.

[11] Hájek,P., Hájková,M.,Havránek, T., Daniel, M. (1999), The Expert System Shell EQUANT-PC: Brief Information. Kybernetika, **25** (1–3) suppl.: 4–9.

[12] Hájek P., Havránek T., Jiroušek R. (1992), *Uncertain Information Processing in Expert Systems*. CRC Press, Boca Raton, Florida.

[13] Hájek P., Valdés (1991), Generalized algebraic foundations of uncertainty processing in rule-based expert systems (dempsteroids). *Computers and Artificial Intelligence* **10** (1): 29–42.

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

[15] Valdés J. J. (1987), *Algebraic and logical foundations of uncertainty processing in rule-based expert systems of Artificial Intelligence.* PhD Thesis, Czechoslovak Academy of Sciences, Prague.

An empirical comparison of popular algorithms for learning gene networks

Vera Djordjilović
Department of Statistical Sciences
University of Padova
djordjilovic@stat.unipd.it

Monica Chiogna
Department of Statistical Sciences
University of Padova
monica@stat.unipd.it

Jiří Vomlel
Institute Of Information Theory and Automation
Czech Academy of Sciences
vomlel@utia.cas.cz

Abstract

In this work, we study the performance of different algorithms for learning gene networks from data. We consider representatives of different structure learning approaches, some of which perform unrestricted searches, such as the PC algorithm and the Gobnilp method and some of which introduce prior information on the structure, such as the K2 algorithm. Competing methods are evaluated both in terms of their predictive accuracy and their ability to reconstruct the true underlying network. A real data application based on an experiment performed by the University of Padova is also considered. We also discuss merits and disadvantages of categorizing gene expression measurements.

1 Introduction

The interest in modelling gene networks has increased in recent years for two reasons. It is a widely accepted stance that a number of disorders and pathologies are associated with subtle changes in gene functioning. Better understanding of the mechanism that governs gene expression is an essential first step towards the development of efficient and highly specific drugs acting on molecular level. In addition to that, technological advances seen in the last two decades drastically reduced experimental costs, which made measurements of biological activity more readily available. This led to a growing body of experimentally obtained knowledge that is stored, in numerous forms, in online public databases. One instance is represented by pathway diagrams, which are elaborate diagrams featuring genes, proteins and other small molecules, showing how they work together to achieve a particular biological effect. From a technical point of view, they are networks and can be represented through a graph where genes and their

connections are, respectively, nodes and edges. Although pathway diagrams represent our up-to-date knowledge of the cellular processes, we can not always assume that derived mathematical graphs will be the optimal structure for statistical modelling. There are a number of reasons to consider them tentative models, see [4], and for this reason structure learning is an important task in genomics setting.

In this empirical comparison, we consider representatives of different structure learning approaches, such as the PC algorithm [8], the Gobnilp method [3] and the K2 algorithm [2]. We perform an extensive simulation study in which we study whether the approaches that include prior information, such as K2, perform better than those that rely on data only. We also look at the impact of discretization. In addition to a simulation study, we consider real data from the *Drosophila Melanogaster* experiment performed by the University of Padova [4]. In this experiment that focused on a WNT signalling pathway in a fruit fly, the expression of 12 genes was measured. Figure 5 shows a DAG derived from a WNT pathway diagram, featuring only genes measured in the experiment.

2 Structure learning algorithms

In this empirical study, we consider a number of variants of the PC algorithm [8], the K2 algorithm [2] and the exact Gobnilp method [3]. Of the examined approaches, the K2 algorithm and all modifications of the K2 algorithm considered here, include the prior information. The prior information is in the form of the topological ordering of the studied genes. In the simulation study, we specify the topological ordering according to the true underlying graph. In the real study, we relied on public databases of biological knowledge. In particular, we used the WNT pathway of the KEGG database to construct a DAG for the set of genes under study, from which we, then, derived a topological ordering. The topological ordering is in general not unique. The consequences of its non-uniqueness will not be discussed here.

To summarize, in this empirical study, we consider the following options.

PC The PC algorithm using χ^2 test of independence at the 5% significance level.

PC20 The PC algorithm using χ^2 test of independence at the 20% significance level.

K2 The original K2 algorithm.

K2-BIC A modified K2 algorithm, where the criterion used to score competing DAGs is BIC, while the search strategy remains the one step greedy search.

G-BIC The Gobnilp algorithm with the BIC scoring criterion.

G-BICm The Gobnilp algorithm with the modified BIC criterion (the penalty term is multiplied by a factor of 10^{-3}).

G-BICl The Gobnilp algorithm where the modified BIC criterion (the penalty term is multiplied by 10^{-9}). This implementation efficiently finds the model with the least number of parameters among all those maximising the log likelihood function.

CK2 The CK2 algorithm proposed in [4]. The only algorithm in this study that is applied to the continuous measurements.

2.1 Categorization of expression measurements

Most structure learning algorithms make use of categorical variables, while gene expressions are quantitative measurements, usually continuous. In the work that first introduced the idea of using DAGs for representing gene regulatory networks, [7] considered both discrete and continuous models. It is clear that the former attenuates the effect of the technical variability, but might lead to information loss, and is sensitive to the choice of the categorization procedure. The former incurs no information loss, but is incapable of capturing non-linear relationships between genes. In particular, combinatorial relationships (one gene is over-expressed only if a subset of its parents is over-expressed, but not if at least one of them is under-expressed) can be modeled only with a discrete Bayesian network. The two approaches thus seem complementary and we believe that both can help researchers obtain the biologically relevant results, at least as a means of postulating testable scientific hypothesis.

When the goal of categorization is to obtain categories which are meaningful from the biological perspective, one would ideally have the control group (a previous experiment) which would serve as a reference for comparison [7]. When control data are not available, we propose to perform categorization based solely on data at hand. It is assumed that genes can assume only a few functional states, for example "under-expressed", "normal", and "over-expressed". The actual measurements depend on these functional states and the amount of biological variability and technical noise. A plausible model for such data is a mixture of K normal distributions, each centered at one of the K functional states

$$X_i \sim \sum_{k=1}^{K} \tau_{ik} \mathsf{N}(\mu_{ik}, \sigma_{ik}^2), \quad i = 1, \ldots, p,$$

where X_i is an expression of the considered gene, μ_{ik} and σ_{ik}^2 are parameters corresponding to the k-th functional state, τ_{ik} the probability that an observation belongs to the k-th component ($\tau_{ik} \geq 0, \sum_{k=1}^{K} \tau_{ik} = 1$) and p is the number of considered genes. However, it is not always plausible to assume that all K states are present in a single experiment, for example, certain genes remain normally expressed in a wide range of conditions, others can only be downregulated, etc. This led us to propose a data driven approach to categorization: a number of components, that can vary from one (corresponding to a gene with only one observed state) to K (all functional states are present in the data) is estimated from the data for each gene independently. The assumed model for the i-th gene is thus

$$X_i \sim \sum_{k=1}^{\hat{K}_i} \tau_{ik} \mathsf{N}(\mu_{ik}, \sigma_{ik}^2), \quad i = 1, 2, \ldots, p,$$

where $\hat{K}_i$ is the estimated number of components for the i-th gene, τ_{ik} are, as before, the weights of individual components, μ_{ik}, σ_{ik} are component specific parameters.

The approach that simultaneously estimates the number of components in the mixture and parameters pertaining to different components and then classifies each observation according to the estimated model is called Model Based Clustering and was introduced by [5]. We used its implementation in the R package `mclust` [6]. In what follows, we will denote $Y_i = (Y_{i1}, \ldots, Y_{i\hat{K}_i})$ the variable obtained from X_i through the proposed categorization, where $Y_{ij} = 1$, if X_i falls to category j, and zero otherwise.

2.2 Evaluation of predictive accuracy

When evaluating the predictive accuracy of different approaches, we restricted our attention to a case with small sample size; a situation most relevant for our field of application. We adopted a "leave-one-out" approach, where in each step the chosen learning algorithm is applied to the data from which the single observation j has been removed. In the second step, the removed observation is used to evaluate the predictive accuracy: prediction of the value of every variable is computed given the values of all other variables.

To measure the distance between the observed value and the predicted value for variable Y_i fixing all remaining variables to the values observed on the removed observation j, we use the Brier score, introduced in [1]. If we denote ${}_jy_i = ({}_jy_{i1}, \ldots, {}_jy_{i\hat{K}_i})$ the observed value of variable Y_i in the jth observation, $j = 1, \ldots, n$, the Brier score is defined as

$$ {}_jb_i = \frac{1}{2} \sum_{k=1}^{\hat{K}_i} ({}_j\hat{\pi}_{ik} - {}_jy_{ik})^2, \tag{1} $$

where ${}_j\hat{\pi}_{ik}$ is the predicted probability that Y_i falls into the category k. The Brier score measures the squared distance between the forecast probability distribution and the observed value. It can assume values between 0 (the perfect forecast) and 1 (the worst possible forecast).

We measure the predictive accuracy with a scalar $B = \sum_{j=1}^{n} \sum_{i=1}^{p} {}_jb_i$. Obviously, algorithms having lower score are preferred.

We compare algorithms designed for categorical and continuous data. The learning algorithms that work with continuous data produce predictions on the continuous scale. In order to make them comparable with categorical predictions, we combine discriminant analysis with the proposed categorization procedure. We classify continuous predictions into one of the gene specific components estimated in the initial categorization. More precisely, we apply the discriminant analysis to the prediction ${}_j\hat{X}_i$; the output is the estimated vector of probabilities $({}_j\hat{\pi}_{i1}, \ldots, {}_j\hat{\pi}_{i\hat{K}_i})$ that ${}_j\hat{X}_i$ falls into associated categories. We can then plug this vector in the expression for the Brier score (1).

3 Simulation study

To attenuate dependence of our conclusions on characteristics of individual graphs, we randomly generated 10 DAGs on 10 nodes. We achieved this by randomly generating

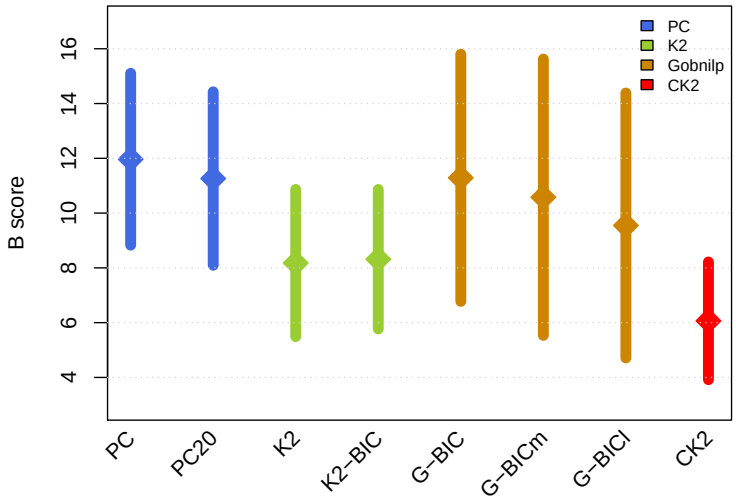


Figure 1: Simulation study: mean value of the B score and its 95% confidence interval.

10 adjacency matrices – for each graph we set a sparsity parameter $\pi \in (0.3, 0.5)$ and fixed the topological ordering. We next sampled an observation from a Bernoulli variable with the parameter π for each plausible edge (corresponding to the upper triangular part of the adjacency matrix) to obtain an adjacency matrix uniquely determining the corresponding DAG. When generating observations from a single DAG, our intention was to mimic the situation in which each gene has two underlying states (low and high expression), that are then affected and, to a certain level, "masked" by some biological and technical variation. We thus generated observations from a mixture of two multivariate normal distributions with a given graphical structure (the so-called Gaussian Bayesian networks, each with weight 0.5), where parameters of each component were randomly sampled from prespecfied intervals. To generate observations for a single component we adopted the structural equations approach, in which each variable is a linear function of its parents and a random error. More precisely, for each of the two components we have

$$X_i = \alpha_i + \boldsymbol{\beta}_i^{\mathsf{T}} \mathrm{pa}(X_i) + \epsilon_i, \quad i = 1, \ldots, p,$$

where $\epsilon_i \sim \mathsf{N}(0, \sigma_i^2)$ is the random disturbance, $\boldsymbol{\beta}_i$ is the vector of regression coefficients giving dependence of X_i on its parents, and α_i is an intercept. Both components were set to have the same matrix of $\boldsymbol{\beta}$ coefficients, so that the dependence structure is shared across components, while the intercept and the random fluctuation were allowed to vary. Before passing these datasets to the algorithms using categorical

variables, we performed categorization as described in 2.1. Namely, we performed model based clustering, where each variable was allowed to have either two or three clusters, depending on the model fit. In this situation, we knew that there were two underlying states–corresponding to two clusters–but we estimated the number of clusters from data so as to approach the conditions of a real study as close as possible. For each graph, we randomly generated 100 datasets.

We first look at the ability of considered algorithms to reconstruct the underlying graphical structure from observations. We rely on two measures: PPV that stands for Positive Predictive Value and is defined as $TP/(TP + FP)$; and Sensitivity, defined as $TP/(TP + FN)$, where TP (true positive), FP (false positive), and FN (false negative) refer to the inferred edges. For each considered sample size and for each of the 10 DAGs, we generated 100 datasets and applied structure learning algorithms. The pooled results are shown in Tables 1 and 2 and Figure 2, that shows graphically how PPV and Sensitivity change with sample size for different approaches. Given that the results of the approaches of the same type (such as PC and PC20; and K2 and K2BIC) have nearly identical results, we show one representative per group, namely PC, GBIC and K2. We see that CK2 gives best results in terms of PPV, and even more strikingly in terms of sensitivity. CK2 is followed by the other two (categorical) K2 approaches and Gobnilp methods. On the other hand, PC algorithm performs poorly in this setting. An interesting question is whether these measures of performance depend on the density of the true underlying DAGs. Figure 3 shows how PPV and Sensitivity depend on the number of edges of the DAG used to generate data. For each of the 10 DAGs, we show the value of PPV and Sensitivity for the largest sample size $n = 500$. We see, perhaps not surprisingly, that PPV increases roughly linearly with the number of edges in the underlying DAG, while sensitivity seems largely unaffected. As an illustration of the performance of considered approaches in reconstructing the "true" DAG, we show one example of a reconstructed network in Figure 4. Alongside a "true" DAG used to simulate data there is a DAG inferred by the CK2 algorithm, from one of the 100 simulated datasets ($n = 500$).

Next, we look at predictive accuracy of considered algorithms. Here, we restricted our attention to the smallest sample size ($n = 20$) for two reasons. It is the situation most relevant to our field of application, where the number of observations is usually limited. Furthermore, it gives us the opportunity to compare obtained results to those in the real application described in Section 4, since the ratio p/n is approximately the same. Therefore, for each of the 10 DAGs and 100 generated datasets of size $n = 20$, we computed the B score following the "leave-one-out" approach, as described in 2.2. In the end, we performed a random effects meta analysis (assuming that the B score is approximately normally distributed) to combine results for different graphs. The mean B score and its 95% confidence interval are shown in Figure 1. CK2 reached the lowest B score, followed by K2 and K2-BIC. Of all Gobnilp methods, the likelihood one G-BICl leads to the lowest B score. PC variants perform slightly worse than Gobnilp variants, but the difference is less pronounced than in network reconstruction.

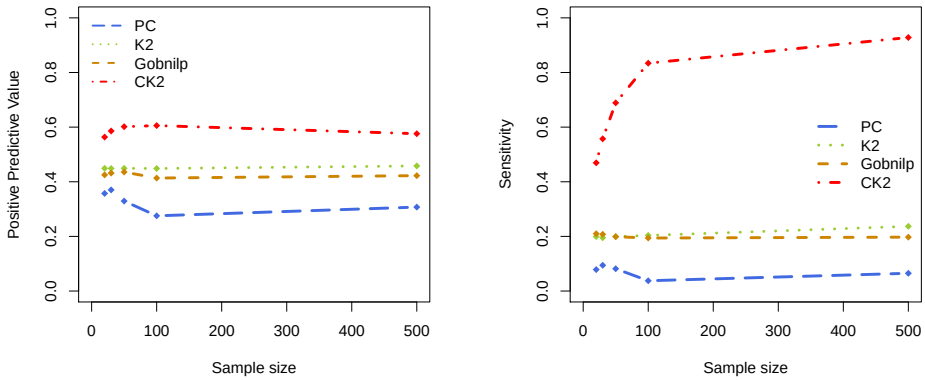


Figure 2: Simulation study: Pooled positive predictive accuracy (left) and sensitivity (right) of considered algorithms for different samples sizes.

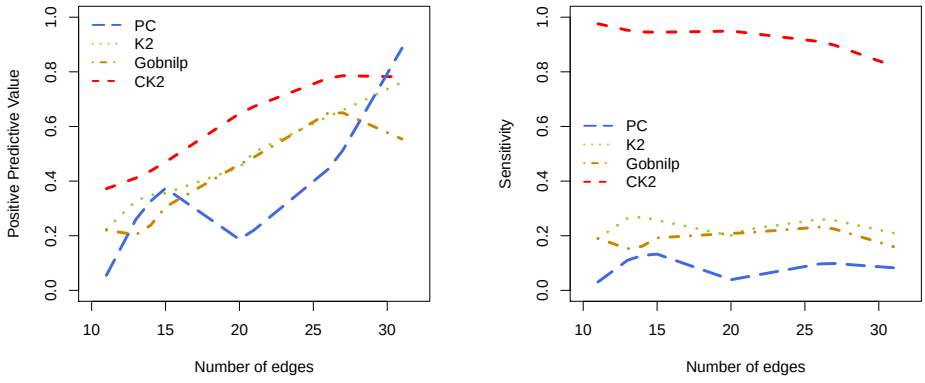


Figure 3: Simulation study: Positive predictive accuracy (left) and sensitivity (right) as a function of the number of edges of the true underlying DAG, for the 10 randomly generated DAGs and the sample size $n = 500$.

Table 1: Pooled positive predictive value.

n	PC	PC20	K2	K2-BIC	GBIC	GBICm	GBICl	CK2
20	0.36	0.35	0.45	0.45	0.43	0.42	0.40	0.56
30	0.37	0.37	0.45	0.45	0.43	0.41	0.40	0.59
50	0.33	0.34	0.45	0.45	0.44	0.41	0.40	0.60
100	0.28	0.25	0.45	0.45	0.41	0.39	0.40	0.61
500	0.31	0.30	0.46	0.46	0.42	0.40	0.41	0.58

Table 2: Pooled sensitivity.

n	PC	PC20	K2	K2-BIC	GBIC	GBICm	GBICl	CK2
20	0.08	0.09	0.20	0.19	0.21	0.21	0.21	0.47
30	0.09	0.10	0.19	0.19	0.21	0.20	0.21	0.56
50	0.08	0.09	0.20	0.20	0.20	0.21	0.20	0.69
100	0.04	0.04	0.20	0.20	0.19	0.21	0.21	0.83
500	0.06	0.07	0.24	0.24	0.20	0.22	0.22	0.93

4 Drosophila Melanogaster experiment

The experimental data from the *Drosophila Melanogaster* experiment performed by the University of Padova [4] consist of 28 observations of 12 genes. All measured genes belong to the WNT signalling pathway involved in embryonic development. DAG derived from this pathway is shown in Figure 5. The topological ordering of this DAG was passed to the methods that include prior information (K2, K2-BIC and CK2). Other methods rely on data only.

The Figure 6 shows the B score for each of the considered methods. Full (complete) DAG and empty (no arrows) DAG were added for reference. Here, K2 reaches the minimal B score, followed by the Gobnilp's likelihood method G-BICl. The K2 algorithm with the BIC score, K2-BIC, together with the remaining Gobnilp methods, G-BICm and G-BIC, also perform reasonably well with a slightly inferior score with respect to the leading twosome. On the other hand, the PC algorithm gives significantly less accurate predictions. The CK2 algorithm, seems to fail in this case. Its B score is almost comparable to the one of the full graph (Full). It is interesting to note that of the two methods on categorized variables using the BIC score, K2-BIC and G-BIC, it is the former that minimizes the B score. This is a little surprising, since Gobnilp finds globally optimal structures, while K2-BIC uses the ordering of variables, and thus might suffer from misspecification. In addition to that, K2-BIC relies on the greedy search, possibly restricting the search space enough to miss the global optima. In fact, structures found by Gobnilp have a lower BIC criterion (and thus a better fit to the data), but are inferior when it comes to prediction. This observation, together with a success of the K2, suggests that possibly the subject matter knowledge employed to specify the ordering of variables is the reason behind their good performance. To test this hypothesis, we generated 20 random orderings and passed them to the K2 algorithm. None of the twenty computed scores is lower than that that determined by pathway, providing support for the practice of using

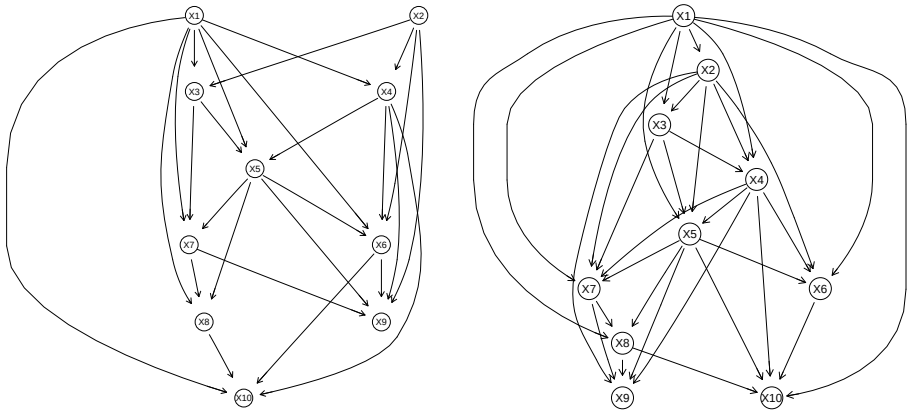


Figure 4: Simulation study: One of the 10 DAGs used to simulate data (left) and the network reconstructed by CK2 from 500 observations.

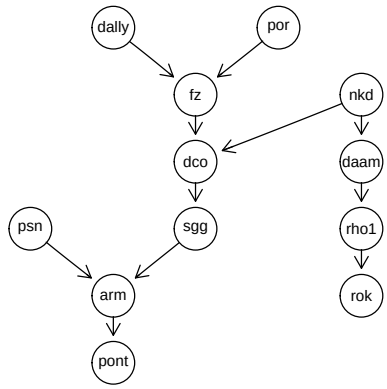


Figure 5: *Drosophila melanogaster* experiment: DAG derived from a diagram representing WNT signaling pathway in fruit flies.

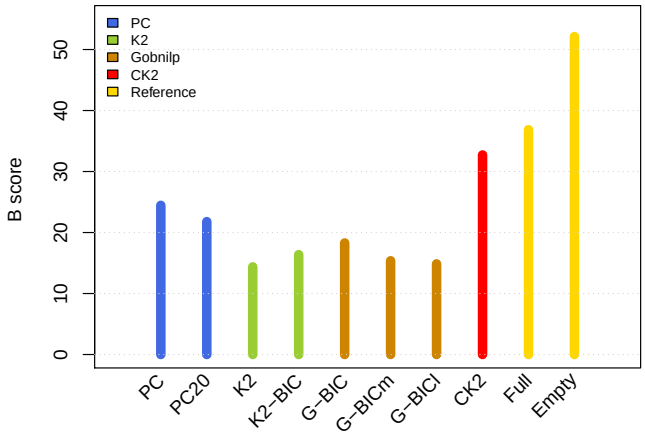


Figure 6: *Drosophila melanogaster* experiment: B score of different algorithms.

the prior information in the form of a topological ordering.

The right plot in Figure 7 shows how the B score deteriorates with the addition of arrows to the optimal structure found by K2. Here, the B score is a function of the number of arrows present in the graph. It starts from the K2 structure, containing 15 arrows, and ends with the full graph, containing 66 arrows. Structures in between are obtained sequentially, by randomly adding a single arrow to the current structure. Obviously, the order of addition of arrows plays a role, and thus this is only one possible way in which the score might evolve between the two extreme points. Nevertheless, the increasing trend of the dependence is informative and independent of the order of arrow inclusion.

One of the reasons behind the success of the K2 algorithm might also be that it identifies DAGs with a relatively high number of edges. To examine this possibility, we computed the average size of the Markov blanket for considered methods. The results are reported in the Table shown in the left panel of Figure 7. We see that K2 indeed has a comparatively large average Markov blanket size, but it is second to the Gobnilp's likelihood method. The ranking of methods with respect to their prediction accuracy suggests therefore that the density of the graphs inferred by K2 is not the only reason for its good performance.

5 Discussion

In this work we performed an extensive empirical study of popular structure learning algorithms in a highly specific setting of gene networks. This area is atypical in that

	Average size
PC	0.98
PC20	1.32
K2	2.64
K2-BIC	2.30
G-BIC	2.15
G-BICm	2.49
G-BICl	2.71
CK2	4.10

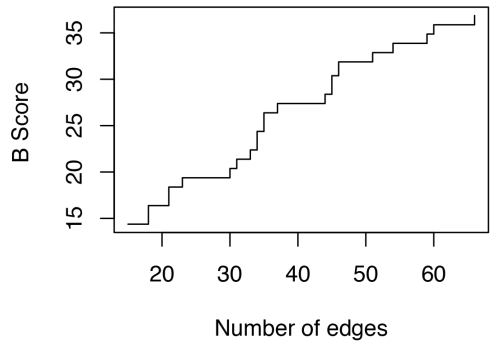


Figure 7: *Drosophila melanogaster* experiment: Average size of the Markov blanket for different algorithms (left) and B score as a function of the number of edges in the inferred DAG.

it usually involves a limited number of observations affected by different kinds of substantial "noise", both biological and technical. For this reason, structure learning in genomics faces a lot of previously unexplored problems and our goal was to better understand the choices made in practice. In particular, we focused on impact of categorising gene expression measurements and including vague prior information. To this end, we analysed a real dataset and performed a simulation study specifically designed to mimic limitations of real studies.

We found that including prior information in the form of a topological ordering can significantly improve the performance, both in terms of network reconstruction and predictive accuracy. This is reflected in the fact that K2 algorithm, in spite of relying on a heuristic search method, performs either better or equally well as the exact Gobnilp method not including any prior information. This observation is especially important with the limited number of observations and was confirmed by both real and simulated datasets.

Results of the simulation study and the real study coincide to a large extent. The most striking difference is the performance of the CK2 algorithm, the only considered algorithm designed for continuous variables. While it performs poorly in the real study, in the simulation study it gives the best results. One possible explanation concerns the simulation mechanism: the data generating mechanism specified in the simulation study might not be a good approximation of the mechanism that gave rise to measurements in the real study. CK2, relying on continuous measurements, would be more sensitive to this difference with respect to its competitors using categorized data. Possible future work would involve investigation of different data generating mechanisms. It would be highly interesting to generate data from a discrete Bayesian network and then introduce random fluctuation for each variable independently.

There is a lot of concern regarding the application of structure learning algorithms in genomics setting. When the goal is to elucidate biological mechanisms governing gene expression, reflected in the reconstruction of the gene network, we would agree

that this concern is justified. The signal to noise ratio in genomic studies does not seem to allow for an accurate reconstruction, at least for the time being. From the prediction perspective, however, the results reported here are encouraging: learned graphs, that can be considered as rough approximations of the true network, manage to bring considerable improvement over the procedure that does not assume or look for any conditional independence relations between genes. This is an important empirical conclusion that we draw from this study.

References

[1] Brier G. W., (1950), Verification of forecasts expressed in terms of probability, *Monthly weather review*, 78(1):1–3.

[2] Cooper, G.F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*. 9(4):309–347.

[3] Cussens, J. and Bartlett, M. (2013). GOBNILP 1.6.2.

[4] Djordjilović, V. (2105). *Graphical modelling of biological pathways*. PhD thesis, University of Padova.

[5] Fraley, C. and Raftery, A.E., (2002), Model-based clustering, discriminant analysis, and density estimation *Journal of the American Statistical Association*, 97(458):611–631.

[6] Fraley C., Raftery A.E., Murphy T.B. and Scrucca L. (2012), `mclust` Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation, *Technical Report*.

[7] Friedman N., Linial M., Nachman I., and Pe'er D. (2000), Using Bayesian networks to analyze expression data, *Journal of computational biology*, 7(3-4):601–620.

[8] Spirtes, P. and Glymour, C.N. and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Utilization of Imprecise Rules Induced by MLEM2 Algorithm

Masahiro Inuiguchi, Takuya Hamakawa
Graduate School of Engineering Science
Osaka University
inuiguti@sys.es.osaka-u.ac.jp

Seiki Ubukata
Graduate School of Engineering
Osaka Prefecture University
subukata@cs.osakafu-u.ac.jp

Abstract

In this paper, we continue the investigation of utilization of rules whose conclusions are imprecise. We examine classification methods under imprecise rules and show their usefulness. Although a set of imprecise rules for all possible combinations of classes improve the classification accuracy, the number of the rules becomes very big. Then we try to reduce the number of imprecise rules keeping the advantage in classification accuracy over the classical approach. Examining a few criteria to select combinations of classes to which rules are induced, we find two tendencie: (i) the set of class combinations with high similarity shows a good performance, and (ii) the set of class combinations reflecting well the distribution of classes in the data set shows a good performance.

1 Introduction

In the conventional rough set approaches, rules inferring the memberships to single classes (simply called "precise rules") have been induced and used to build the classifier system. However, rules inferring the memberships to unions of multiple classes (simply called "imprecise rules") can also be induced based on the rough set model. We have shown that a classifier system with imprecise rules has an advantage in the classification accuracy of classification over the conventional classifier system with precise rules [1, 2]. However, the number of imprecise rules is much more than that of precise rules because we consider all possible combinations of k classes with fixed $k \in (1, p-1)$.

In this paper, we try to reduce the number of imprecise rules keeping the classification accuracy of the classifier with imprecise rules over that of the classifier with precise rules. We first propose a new classification method using imprecise rules and demonstrate its advantage over the previous classification method. To reduce the number of imprecise rules, we select combinations of classes. Because of its simplicity, we restrict ourselves to the cases where only two classes are combined. Namely, we investigate some criteria (measures) for selecting class pairs. We consider two conceivable criteria for the selection: one is the similarity of classes and the other is

the deviation from the class distribution. For the similarity, we would like to know whether similar classes should be paired or dissimilar classes in order to enhance the classification accuracy. On the other hand, for the deviation from the class distribution, we examine whether selected class pairs should reflect the class distribution in the given data set or not in order to enhance the classification accuracy.

In next section, we briefly review rough set approach and a rule induction method, i.e., MLEM2 algorithm, and the conventional classification method based on the induced rules. In Section 3, after the idea of inducing imprecise rules is described, two classification methods based on induced imprecise rules are proposed. Moreover, the similarity and KL divergence are introduced as measures to select the set of class combinations. In Section 4, numerical experiments are explained and the results are shown and discussed. In Section 5, concluding remarks are given together with future research topics.

2 Rough Set Approach and Rule Induction

Rough set theory [3] provides useful tools for the analysis of decision tables which is also called datasets. A decision table (dataset) is defined by a four-tuple $DT = \langle U, C \cup \{d\}, V, f\rangle$, where U is a finite set of objects, C is a finite set of condition attributes and d is a decision attribute, $V = \bigcup_{a \in C \cup \{d\}} V_a$ with attribute value set V_a of attribute $a \in C \cup \{d\}$ and $f : U \times C \cup \{d\} \to V$ is called an information function which is a total function. By decision attribute value $v_j^d \in V_d$, decision class $D_j \subseteq U$ is defined by $D_j = \{u \in U \mid f(u, d) = v_j^d\}$, $j = 1, 2, \ldots, p$. Using condition attributes in $A \subseteq C$, we define equivalence classes $[u]_A = \{x \in U \mid f(x, a) = f(u, a),\ \forall a \in A\}$.

The lower and upper approximations of an object set $X \subseteq U$ under condition attribute set $A \subseteq C$ are defined by

$$A_*(X) = \{x \in U \mid [x]_A \subseteq X\}, \tag{1}$$

$$A^*(X) = \{x \in U \mid [x]_A \cap X \neq \emptyset\}. \tag{2}$$

Suppose that members of X can be described by condition attributes in A. If $[x]_A \cap X \neq \emptyset$ and $[x]_A \cap (U - X) \neq \emptyset$ hold, the membership of x to X or $U - X$ is questionable because objects described in the same way are classified into two different classes. Otherwise, the classification is consistent. From these points of view, each element of $A_*(X)$ can be seen as a consistent member of X while each element of $A^*(X)$ can be seen as a possible member of X. The pair $(A_*(X), A^*(X))$ is called the rough set of X under $A \subseteq C$.

In rough set approaches, the attribute reduction, i.e., the minimal attribute set $A \subseteq C$ satisfying $A_*(D_j) = C_*(D_j)$, $j = 1, 2, \ldots, p$, and the minimal rule induction, i.e., inducing rules inferring the membership to D_j with minimal conditions which can differ members of $C_*(D_j)$ from non-members, are investigated well. In this paper, we use minimal rule induction algorithms proposed in the field of rough sets, i.e., LEM2 and MLEM2 algorithms [4]. By those algorithms, we obtain minimal set of rules with minimal conditions which can explain all objects in lower approximations of X of the given dataset. LEM2 algorithm and MLEM2 algorithm [4] are different in their forms

of condition parts of rules: by LEM2 algorithm, we obtain rules of the form of "if $f(u, a_1) = v_1$, $f(u, a_2) = v_2$, ... and $f(u, a_p) = v_p$ then $u \in X$", while by MLEM2 algorithm, we obtain rules of the form of "if $v_1^{\mathrm{L}} \leq f(u, a_1) \leq v_1^{\mathrm{R}}$, $v_2^{\mathrm{L}} \leq f(u, a_2) \leq v_2^{\mathrm{R}}$, ... and $v_p^{\mathrm{L}} \leq f(u, a_p) \leq v_p^{\mathrm{R}}$ then $u \in X$". Namely, MLEM2 algorithm is a generalized version of LEM2 algorithm to cope with numerical/ordinal condition attributes.

For each class D_i we induce rules inferring the membership of D_i. Using all those rules, we build a classifier system. To build the classifier system, we apply the idea of LERS [4]. The classification of a new object u is made by the following two steps:

1. When the condition attribute values of u match to at least one of the elementary conditions of the rule, we calculate

$$S(D_i) = \sum_{\text{matching rules } r \text{ for } D_i} Stren(r) \times Spec(r), \tag{3}$$

where r is called a *matching rule* if the condition part of r is satisfied. $Stren(r)$ is the total number of objects in the given dataset correctly classified by rule r. $Spec(r)$ is the total number of condition attributes in the condition part of rule r. For convenience, when rules from a particular class D_i are not matched by the object, we define $S(D_i) = 0$. If there exists D_j such that $S(D_j) > 0$, the class D_i with the largest $S(D_i)$ is selected. If a tie occurs, class D_i with smallest index i is selected from tied classes.

2. When the condition attribute values of u do not match totally to any condition part of rule composing the classifier system. For each D_i, we calculate

$$M(D_i) = \sum_{\substack{\text{partially matching} \\ \text{rules } r \text{ for } D_i}} Mat\_f(r) \times Stren(r) \times Spec(r), \tag{4}$$

where r is called a *partially matching rule* if a part of the premise of r is satisfied. $Mat\_f(r)$ is the ratio of the number of matched conditions of rule r to the total number of conditions of rule r. Then the class D_i with the largest $M(D_i)$ is selected. If a tie occurs, class D_i with smallest index i is selected from tied classes.

3 Imprecise Rules and Evaluation Measures

3.1 Induction of Imprecise Rules and Classification

In the same way as inducing rules inferring the membership to D_i, we can induce rules inferring the membership to the union of D_i's. Namely, LEM2-based algorithms can be applied because the union of D_i's is a set of objects. Inducing rules inferring the membership of the union of $D_i \cup D_j$ for all pairs (D_i, D_j) such that $i \neq j$, we may build a classifier because the simultaneous satisfaction of $x \in D_i \cup D_j$ and $x \in D_i \cup D_k$ $(j \neq k)$ mplies $x \in D_i$. Moreover, in the same way, we can build a classifier by induced rules inferring the membership to $\bigcup_{j=i_1,i_2,\ldots,i_l} D_j$ for all combinations of l classes.

To do this, we should consider a classification method under rules inferring the membership to the union of classes. An easiest method is using the MLEM2 classification method described in the previous section with replacing $S(D_i)$ and $M(D_i)$ with the following $\hat{S}(D_i)$ and $\hat{M}(D_i)$, respectively:

$$\hat{S}(D_i) = \sum_{\substack{\text{matching rule } r \\ \text{for } Z \supseteq D_i}} Stren(r) \times Spec(r), \tag{5}$$

$$\hat{M}(D_i) = \sum_{\substack{\text{partially matching} \\ \text{rules } r \text{ for } Z \supseteq D_i}} Mat\_f(r) \times Stren(r) \times Spec(r), \tag{6}$$

where Z is a variable showing a union of classes.

Obviously, this classification method is reduced to the LERS classification method when $Z = D_i$. For the sake of simplicity, this method is called CL-1.

As another one, we propose the following classification method:

(i) we calculate $S(Z_q)$ by (3) with substitution of Z_q for D_i for each $Z_q = \bigcup_{j \in \{i_1, i_2, \ldots, i_l\}} D_j$.

(ii) For each Z_q such that $S(Z_q) = 0$, erase all D_j satisfying $D_j \subseteq Z_q$.

(iii) If the remaining class is unique, then we classify the object into that class and terminate the procedure. If the remaining class is empty, we reset all classes as remaining classes.

(iv) For each remaining class D_i, calculate $\hat{S}(D_i)$ by (5).

(v) Classes D_i with the largest $\hat{S}(D_i)$ are selected. If it is unique, then we classify the object into that class and terminate the procedure. Otherwise, for each remaining class D_i, calculate $\hat{M}(D_i)$ by (6), and class D_i with the largest $\hat{M}(D_i)$ is selected. If a tie occurs, class D_i with smallest index i is selected from tied classes.

This method is called CL-2 in this paper.

3.2 Measures

3.2.1 Similarity between combined classes

We use this similarity measure among classes defined by Kusunoki and Inuiguchi [5]. In numerical experiments, we examine whether the similarity of combined classes in imprecise rules influences the classification accuracy of the classifier composed. The similarity is computed in the following two steps:

i) for each attribute $a \in C$, the similarity between two objects $x, y \in U$ is defined by

$$s(x, y; a) = \begin{cases} \text{Truth}(f(x, a) = f(y, a)), \\ \qquad \text{if } a \text{ is nominal}, \\ \max\left\{1 - \dfrac{|x - y|}{3\sigma_a}, 0\right\}, \\ \qquad \text{if } a \text{ is numerical}, \end{cases} \tag{7}$$

where Truth(*proposition*) takes 1 if *proposition* is true and 0 otherwise; σ_a is the standard deviation of values of condition attribute a in the dataset. The similarity between objects x and y is defined as an arithmetic mean of similarities $s(x, y; a)$ over all attributes $a \in C$, i.e.,

$$s(x, y) = \frac{1}{|C|} \sum_{a \in C} s(x, y; a), \tag{8}$$

where $|C|$ is the cardinality of set C.

ii) Then, the similarity between two classes $X, Y \subseteq U$ is defined by

$$s(X, Y) = \frac{1}{|B(X)||B(Y)|} \sum_{x \in B(X)} \sum_{y \in B(Y)} s(x, y), \tag{9}$$

where $B(X)$ and $B(Y)$ are $A_*(X)$ and $A^*(Y)$, respectively.

3.2.2 Distance from the class distribution

In a dataset, the sizes of classes can be different. We may intuitively expect that the obtained set of rules performs better when the selected unions of classes reflect the class distribution. From this point of view, we consider the distance from the class distribution.

Let O_i, $i = 1, 2, \ldots, u$ be selected unions O_i of classes to which we induce rules. Then we define the following appearance $ap(D_j)$ for each class D_j $(j = 1, 2, \ldots, p)$:

$$ap(D_j) = |\{O_i \mid D_j \subseteq O_i,\ i \in \{1, 2, \ldots, u\}\}|. \tag{10}$$

Then we define the relative appearance $ra(D_j)$ by

$$ra(D_j) = \frac{ap(D_j)}{\sum_{j=1}^{p} ap(D_j)}. \tag{11}$$

Let rf_j be the relative frequency of class D_j in the given dataset $(j = 1, 2, \ldots, p)$. Then the distance from the class distribution is defined by the following KL divergence from the relative frequency distribution to the relative appearance distribution;

$$KL(\mathcal{O}) = \sum_{j=1}^{p} rf_j \log \frac{rf_j}{ap(D_j)}, \tag{12}$$

where $\mathcal{O} = \{O_i, i = 1, 2, \ldots, u\}$.

Table 1: Five datasets

Dataset	$\|U\|$	$\|C\|$	$\|V_d\|$	attribute type	class
car (C)	1,728	6	4	ordinal	umacc, acc, good, vgood
dermatology (D)	358	34	6	numerical	1, 3, 2, 5, 4, 6
ecoli (E)	336	7	8	numerical	cp, im, pp, imS, om, omL, imU, imL
glass (G)	214	9	6	numerical	2, 1, 7, 3, 5, 6
zoo (Z)	101	16	7	nominal	1, 2, 4, 7, 6, 3, 5

4 Numerical Experiments

4.1 Outline

We first demonstrate the good performance of the classifier based on the imprecise rules obtained for all combinations of classes by a numerical experiment using five datasets. Then we reduce the number of rules by restricting ourselves into cases when two classes are combined. For those rules, we compute the number of rules, the classification accuracy, the KL divergence and the similarity, in order to see their relations. We evaluate the classification accuracy by both CL-1 and CL-2.

4.2 Datasets

In the numerical experiments, we use five datasets obtained from UCI Machine Learning Repository [6]. The five datasets are shown in Table 1. In Table 1, $|U|$, $|C|$ and $|V_d|$ means the number of objects in the given data table, the number of condition attributes and the number of classes. Decision attribute values are shown in the column of 'class' and labeled by alphabets a, b, c, ... in the order of attribute values shown in Table 1. The capital alphabets in the parentheses in the column of dataset shows the abbreviations of the dataset names.

MLEM2 algorithm is applied to all those datasets because MLEM2 algorithm produces the same results as LEM2 when all condition attributes are nominal.

4.3 10-fold Cross Validation

For the evaluation, we apply the 10-fold cross validation method. Namely we divide the dataset into 10 subsets and 9 subsets are used for training dataset and the remaining subset is used for checking dataset. Changing the combination of 9 subsets, we obtain 10 different evaluations. We calculate the averages and the standard deviations in each evaluation measure. We execute this procedure 10 times with different divisions.

4.4 Accuracy Scores of Imprecise Rules

First, we demonstrate the good performance of the classifier with imprecise rules. Using training data, we induce rules inferring the membership to the union of k classes by MLEM2 algorithm for all possible combination of k classes. Then by the

Table 2: Accuracy scores of imprecise rules for all class combinations

$A(k)$	No. Rules	CL-1 (%)	CL-2 (%)
C(1)	57.22 ± 1.74	98.67 ± 0.97	98.67 ± 0.97
C(2)	128.02 ± 3.16	98.96 ± 0.73	$\underline{99.16} \pm 0.76$
C(3)	69.55 ± 1.37	$\underline{99.68} \pm 0.49$	99.57 ± 0.54
D(1)	12.09 ± 1.27	92.32 ± 4.42	92.32 ± 4.42
D(2)	61.32 ± 4.07	94.58 ± 3.59	$\underline{95.72}^* \pm 3.15$
D(3)	103.58 ± 6.11	96.03 ± 3.26	$\underline{96.40} \pm 3.14$
D(4)	77.28 ± 4.45	95.58 ± 3.69	$\underline{95.78} \pm 3.67$
D(5)	23.84 ± 1.81	$\underline{91.87}^* \pm 4.75$	88.83 ± 5.73
E(1)	35.89 ± 2.03	75.52 ± 6.21	75.52 ± 6.21
E(2)	220.67 ± 8.93	83.20 ± 5.66	$\underline{83.42} \pm 5.56$
E(3)	565.67 ± 21.48	$\underline{84.66} \pm 5.64$	84.54 ± 5.75
E(4)	781.36 ± 28.42	$\underline{84.87} \pm 5.71$	84.84 ± 5.65
E(5)	617.06 ± 23.06	$\underline{83.74} \pm 6.26$	83.53 ± 6.38
E(6)	269.27 ± 10.50	82.56 ± 6.26	$\underline{82.76} \pm 6.27$
E(7)	54.09 ± 2.86	$\underline{78.38} \pm 6.70$	77.17 ± 6.71
G(1)	25.38 ± 1.50	63.34 ± 10.18	68.34 ± 10.18
G(2)	111.4 ± 4.33	72.57 ± 8.81	$\underline{73.59} \pm 8.77$
G(3)	178.35 ± 5.41	73.44 ± 9.19	$\underline{74.28} \pm 9.93$
G(4)	130.14 ± 4.96	71.16 ± 9.91	$\underline{72.71} \pm 9.45$
G(5)	39.59 ± 2.18	$\underline{65.04} \pm 9.96$	63.55 ± 10.79
Z(1)	9.67 ± 0.55	95.84 ± 6.63	95.84 ± 6.63
Z(2)	48.54 ± 2.10	95.55 ± 7.15	$\underline{95.74} \pm 6.33$
Z(3)	105.37 ± 4.25	96.74 ± 5.45	96.74 ± 5.45
Z(4)	113.78 ± 3.74	96.84 ± 5.22	96.84 ± 5.22
Z(5)	66.76 ± 2.69	97.24 ± 5.07	$\underline{97.44} \pm 4.97$
Z(6)	17.72 ± 0.66	96.05 ± 6.51	96.05 ± 6.51

two classifiers based on induced imprecise rules, all of checking data are classified and the classification accuracy scores are calculated. The results of this numerical experiment are shown in Table 2. In Table 2, the average number of induced rules, the average classification accuracy scores by CL-1 and by CL-2 are shown. Column $A(k)$ indicates the abbreviation of dataset by A

Table 3: Three sets of class pairs for each dataset

A	set of pairs			A	set of pairs			A	set of pairs			A	set of pairs		
	i	ii	iii		i	ii	iii		i	ii	iii		i	ii	iii
C	ab	ab	ac	E	ah	ab	ah	G	ab	bd	bf	Z	ab	ac	ae
	ad	ac	ad		ac	ae	af		bd	bf	bc		ad	ag	ad
	bc	bd	bc		bg	bc	be		ad	ad	ad		be	bg	bc
	cd	cd	bd		bd	gh	bf		ef	ac	af		cf	be	bg
D	ad	ab	ac		dg	dg	dg		ec	ef	de		fg	fe	fg
	af	ad	ab		fh	dh	cg		cf	ec	ec		cg	fd	fd
	bc	ce	cd		ef	ef	eh						de	cd	ce
	ce	cd	bf		ec	cf	cd								
	bd	bf	de												
	ef	ef	ef												

Table 4: Relations of classification accuracy to similarity

car	C[i]	C[ii]	C[iii]
No. rules	75.93 ± 2.21	84.97 ± 2.31	95.14 ± 2.48
Accuracy (%)	$\underline{98.48} \pm 0.83$	98.44 ± 1.04	98.45 ± 1.06
Similarity	$\underline{32.45}^{*23} \pm 0.07$	$31.90^{*3} \pm 0.07$	30.52 ± 0.05
dermatology	D[i]	D[ii]	D[iii]
No. rules	20.31 ± 1.39	19.46 ± 1.38	27.90 ± 2.44
Accuracy (%)	$\underline{92.46}^{*3} \pm 4.16$	$92.20^{*3} \pm 4.88$	90.31 ± 4.98
Similarity	$\underline{71.67}^{*23} \pm 7.47$	$70.77^{*3} \pm 9.20$	70.05 ± 9.13
ecoli	E[i]	E[ii]	E[iii]
No. rules	47.28 ± 2.35	65.10 ± 3.40	70.45 ± 3.44
Accuracy (%)	$\underline{80.53}^{*3} \pm 6.59$	78.94 ± 6.70	78.35 ± 6.44
Similarity	$\underline{70.48}^{*23} \pm 4.70$	$64.67^{*3} \pm 3.84$	60.85 ± 4.74
glass	G[i]	G[ii]	G[iii]
No. rules	35.95 ± 1.85	43.44 ± 2.26	48.09 ± 2.74
Accuracy (%)	$\underline{65.62} \pm 11.10$	64.55 ± 9.41	64.55 ± 10.95
Similarity	$\underline{69.85}^{*23} \pm 0.39$	$68.66^{*3} \pm 0.48$	64.06 ± 0.48
zoo	Z[i]	Z[ii]	Z[iii]
No. rules	11.77 ± 0.53	16.72 ± 0.51	15.94 ± 0.40
Accuracy (%)	81.27 ± 11.35	$95.35^{*1} \pm 7.57$	$\underline{96.04}^{*1} \pm 7.04$
Similarity	$\underline{62.08}^{*23} \pm 0.51$	$58.02^{*3} \pm 0.51$	53.54 ± 0.48

and the number of classes composing the union by k. Each entry in the other columns is composed of the average av and the standard deviation st in the form of $av \pm st$. Because we have two classification accuracy scores for each case, we compare those, namely, we compare two classification methods CL-1 and CL-2. Larger average of classification accuracy scores between CL-1 and Cl-2 is underlined. Moreover, asterisk * is attached to the better classification accuracy if there is a significant difference by the paired t-test with significance level $\alpha = 0.05$.

As shown in Table 2, the classification accuracy sometimes attains its largest value

around the middle value of k. We observe that the classifier with imprecise rules often performs better than the classifier with precise rules (when $k = 1$). Moreover, we observe that classification method CL-2 is slightly better than CL-1.

4.5 Relations to Similarity

In the previous subsection, we showed the good performance of the classifier with imprecise rules for all possible combination of k classes except a few datasets with $k = n$. However, the number of imprecise rules is much more than that of precise rules. Therefore, imprecise rules are not always very efficient in rule induction and rule applications. We examine whether the good performances are kept by reducing the number of rules by restricting combinations of k classes.

In this subsection, we examine whether similarity degree works to select a set of class combinations to which we can induce imprecise rules of good performance. In the numerical experiment, we set $k = 2$ and we prepare three sets of class pairs for each dataset. The number of pairs in each set is set to be the minimum subject to the induced rules can correctly classify all given objects in training dataset. For example, when there are four classes, the minimally requested combinations is four while the number of all possible combinations is six. This is because each class must belong to at least two different combinations. Three sets of class pairs for each dataset are shown in Table 3. The columns of "A" in Table 3 show abbreviated names of datasets. We note that the KL divergence values are the same in the three sets of pairs.

Through this experiment, we would like to seek good pairs for obtaining a smaller set of imprecise rules which performs well. For this purpose we compute the similarity degrees of classes as well as numbers of rules. The obtained results are shown in Table 4.

Numerical results are shown in Table 4. In rows where full names of datasets are written, settings are shown by the abbreviations of dataset names with the set of class pairs in the square brackets. A generic entry of this table is composed of the average av and the standard deviation st in the form of $av \pm st$. Mark ^{*23} means the value is significantly different from the cases of sets ii and iii by the paired t-test with significance level $\alpha = 0.05$ (we mark only better values). In this experiment, classification accuracy scores of CL-1 and CL-2 become same, then we only show them in rows of "Accuracy". The largest classification accuracy score and largest similarity degree are underlined in each dataset.

As shown in Table 4, the numbers of rules are reduced from those in the case of $k = 2$ in Table 2. Nevertheless, the numbers of rules are more than those in the case of $k = 1$ in Table 2. All classification accuracy scores in Table 4 are worse than those of $k = 2$ in Table 2 except Z[iii]. However, classification accuracy scores of D[i], E[i], E[ii], E[iii] and Z[iii] are better than those of $k = 1$ in Table 2. Except for dataset 'zoo', the larger the similarity degree, the better the classification accuracy score. Condition attributes in dataset 'zoo' are nominal, while those in other datasets are numerical or ordinal. This may be caused by the fact dataset 'zoo' shows the opposite tendency to the other datasets. Therefore, the similarity degree may be a

measure to select a set of class pairs (class combinations) to induce imprecise rules with good performance if condition attributes are numerical or ordinal.

4.6 Relations to KL Divergence

In this section, we examine whether KL divergence from the class distribution works well to select a set of class combinations to which we can induce imprecise rules of good performance. In the previous experiment, sets of minimal number of class pairs are considered. In the experiment described in this subsection, a few sets with different numbers of class pairs are considered. Here again we consider only imprecise rules with $k = 2$. For each number of class pairs, we prepare two sets of class pairs. Those two set of class pairs take a very different KL divergence values. Comparing classification accuracy scores between those two sets, we observe the influence of KL divergence values.

Three different numbers of class pairs are prepared except dataset 'car'. Because dataset 'car' has four classes, we consider sets of five class pairs only. For each of those sets of class pairs, we calculate the number of rules, classification accuracy scores and KL divergence value. The results of this numerical experiment is shown in Table 5. To show a setting composed of the dataset, the number of class pairs and two sets of different KL divergence values, we use a combined notation of the abbreviation of dataset name, a number and 'L (large) or S (small)' in Table 5. For example, 'D[13-L]' implies dataset 'dermatology' and 13 class pairs with larger KL divergence value. Similarly, 'G[11-S]' implies dataset 'glass' and 11 class pairs with smaller KL divergence value. In the column of KL divergence, we show the class pairs missing in the set of class pairs in the parentheses. Those with '+' shows the additionally missing class pairs from the set of class pairs described two rows above in Table 5. In columns of 'CL-1' and 'CL-2' of Table 5, we show the average (av) and standard deviation (st) of classification accuracy scores for each setting in the form of $av \pm st$. The average and standard deviation of numbers of rules is shown in the column of 'No. rules'. We compare two settings with large (L) and small (S) KL divergence values, and underline the better average of classification accuracy scores. Asterisk * means the value is significantly better from the other by the paired t-test with significance level $\alpha = 0.05$.

As shown in Table 5, all classification accuracy scores in Table 5 are worse than those of $k = 2$ in Table 2 except Z[15-S]. However, classification accuracy scores of C[5-2], D[13-L], D[13-S], D[11-L], D[9-S], Z[18-S], Z[15-S] and all settings in datasets 'ecoli' and 'glass' are better than those of $k = 1$ in Table 2. In many cases, we observe that the smaller KL divergence values, the better the classification accuracy scores. In dataset 'zoo' and in the setting of 9 class pairs of dataset 'dermatology', we obtained opposite results. However, in the case of dataset 'zoo', there is no significance difference in classification accuracy scores. On the contrary, the result in the setting of 9 class pairs of dataset 'dermatology' significantly opposes to the above observation. To sum up, there is some tendency that the set of class pairs with smaller KL divergence score may result in a better classification accuracy score.

Table 5: Comparison between class pair sets with different KL divergence values

Setting	No. rules	CL-1 (%)	CL-2 (%)	KL divergence
C[5-L]	111.59±2.64	98.84*±0.80	98.85*±0.77	0.5766(cd)
C[5-S]	111.57±2.90	98.22±1.11	98.07±1.09	1.0708(ab)
D[13-L]	53.26±3.74	94.56±3.66	94.64±3.58	0.0750(df,ef)
D[13-S]	52.57±3.59	94.38±3.55	95.03±3.12	0.2884(ab,ac)
D[11-L]	43.27±2.99	94.02*±3.75	94.02*±3.60	0.0342(+cf,de)
D[11-S]	44.66±2.95	92.23±4.02	92.71±3.93	0.4230(+ad,bc)
D[9-L]	32.35±1.81	90.45±4.43	90.45±4.43	0.0578(+bc,cd)
D[9-S]	33.57±2.37	92.54*±3.90	92.93*±3.77	0.4062(+bd,de)
E[25-L]	215.73±8.91	83.17±5.64	83.18±5.59	0.6793(fg,fh,gh)
E[25-S]	186.17±7.54	82.97±5.76	82.90±5.83	1.0912(ab,ac,ad)
E[21-L]	193.74±8.29	82.81±5.56	82.78±5.54	0.5288(+cf,dh,eg,eh)
E[21-S]	140.55±5.71	81.84±6.12	81.84±6.12	1.2494(+ae,bc,bf,de)
E[17-L]	158.84±7.87	82.93*±5.38	82.87*±5.39	0.4237(+bg,bh,cg,df)
E[17-S]	100.90±4.19	80.79±6.64	80.79±6.64	1.3943(+af,be,cd,cg)
G[13-L]	103.43±4.12	72.16*±9.07	72.21*±8.56	0.2781(df,ef)
G[13-S]	90.84±3.80	68.14±10.60	68.24±10.78	0.6126(ab,ac)
G[11-L]	95.52±3.88	71.19*±9.02	71.20*±8.69	0.1635(+cf,de)
G[11-S]	70.03±3.10	61.91±10.83	62.05±10.84	0.7969(+ad,bc)
G[9-L]	85.89±3.78	69.42*±9.68	69.47*±9.48	0.1175(+bd,ce)
G[9-S]	54.20±2.58	59.73±11.88	59.73±11.88	0.8306(+be,cd)
Z[18-L]	42.54±2.08	95.25±7.44	95.25±7.44	0.3418(df,eg,fg)
Z[18-S]	40.65±1.98	95.95±6.67	95.54±6.81	0.7607(ab,ac,ad)
Z[15-L]	35.84±1.64	95.16±7.55	95.16±7.55	0.2425(+cg,ef,cd)
Z[15-S]	32.69±1.97	95.95±6.37	96.14±5.95	0.9693(+bc,ac,bd)
Z[12-L]	30.27±1.21	95.15±7.58	95.15±7.58	0.1533(+bg,bf,de)
Z[12-S]	25.79±1.70	95.15±6.89	95.15±6.89	0.7976(+cf,eg,df)

5 Concluding Remarks

In this paper, we first demonstrated the good performance of the classifiers with imprecise by comparing them to the classifiers with precise rules. However, the number of imprecise rules is much more than that of precise rules. Then we tried to reduce the number of imprecise rules keeping its high classification accuracy. To select class combinations to which we induce imprecise rules, we proposed two measures, similarity and KL divergence to class distribution. By the numerical experiments restricting ourselves into pairing classes, we could not observe any strong relations but two tendencies: (i) it would be better to select similar classes than dissimilar classes and (ii) it would be better to select class pairs so that the selected pairs of classes match the class distribution, i.e., small KL divergence would be better.

Except a few cases, in the numerical experiments performed in Subsections 4.5 and

4.6, we could not obtain better classification accuracy scores than the case of $k = 2$ in Table 2. This fact implies that it is very difficult to keep the high classification accuracy score against reducing the number of induced rules. However, we could obtain better classification accuracy scores than the classification scores by the usual MLEM2 rules (the case of $k = 1$ in Table 2) in the numerical experiments performed in Subsections 4.5 and 4.6.

We should continue the investigation for the improvement of the usage of the imprecise rules. For example, we should execute a similar experiments without restricting ourselves into pairing classes and examine the selection by KL divergence value together with similarity degree. Moreover, datasets like 'zoo', the tendencies were not observed. Then we should further investigate the properties of datasets which show the observed tendencies.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 26350423.

References

[1] Inuiguchi, M., Hamakawa, T. (2013), The utilities of imprecise rules and redundant rules for classifiers, in: Huynh, V.-N. et al. (eds.), *Knowledge and Systems Engineering: Proceedings of the Fifth International Conference KSE 2013*, Vol.2, AISC 245, Springer, 45–56

[2] Hamakawa, T, Inuiguchi, M. (2014), On the utility of imprecise rules induced by MLEM2 in classification, in: *Proceedings of 2014 IEEE International Conference on Granular Computing*CIEEE Xplore, 76–81

[3] Pawlak, Z.(1982), Rough sets, *International Journal of Computer and Information Sciences*, vol.11, no.5, 341–356

[4] Grzymala-Busse, J. W. (2003), MLEM2 - Discretization during rule induction, in: *Proceedings of the IIPWM2003*, 499–508

[5] Kusunoki, Y , Inuiguchi, M. (2008), Rule induction via clustering decision classes, *International Journal of Innovative Computing, Information and Control*, vol.4, no.10, 2663–2677

[6] UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/`

Learning correction and turning rules from data

Jiří Ivánek
Department of Information and Knowledge Engineering
University of Economics, Prague
e-mail: ivanek@vse.cz

Abstract

The task of the rule set acquisition from data consists in the question which of empirical implications (association rules) existing in data are to be inserted into the resulting rule set, and with which weights. Our algorithm based on correction principle generates and tests implications in the sequence according to the complexity or frequency of the left-hand side. If the confidence of the implication in data significantly differs from the composed weight (value obtained when composing weights of all sub-rules of the implication which have been inserted into the rule base already) then this implication is added to the rule base with a weight correcting the composed weight to the confidence of the implication in data.

In the contribution, we discuss possible using Łukasiewicz's composition function in this method. We focus on situations when composed weights and confidences are strongly opposite. We solve this by adding to the rule set not only correcting rules, but also special turning rules. Obtained results serve as a basis for modifications of the algorithm for the automatic rule set acquisition from categorical data using Łukasiewicz's composition function.

1 Introduction

The task of the rule base acquisition from data consists in the question which of empirical implications (association rules) existing in data are to be inserted into the resulting rule set, and with which weights. The (one-level) rule set contains for each goal (class) C a set of weighted rules leading from combinations of values of input attributes to this goal. Rules are given in the form of implications

$$[A_1 \wedge \ldots \wedge A_k \Longrightarrow C;\ r] \tag{1}$$

where A_i are standing for values of different attributes, and r is a *weight* (degree). The set of rules are used by some inference mechanism for uncertainty processing in rule-based systems to obtain estimations of probabilities of C for every combinations of values of input attributes. Our aim is to pick up from data as small as possible

number of "important" rules satisfying the request of the accordance of the resulting estimations with the data for those combinations of values of input attributes which are existing in the data. We hope these rules are "pieces of knowledge" derived from data and we would like to eventually combine them with expert rules. For this reason, we try to apply the inference mechanism based on Łukasiewicz's fuzzy logic with evaluated syntax $Ev_{\mathbf{L}}$ (see [6, 10]) which has some advantages, namely connected to the simplicity of combining uncertainties using only operations of limited addition and subtraction of weights. All computations with weights are so clearly understandable for experts.

At first, in Section 2, our general method for rule set acquisition from data is recalled. An algorithm based on correction principle generates and tests implications in the sequence according to the complexity or frequency of the left-hand side. If the confidence of the implication in data significantly differs from the composed weight (value obtained when composing weights of all sub-rules of the implication which have been inserted into the rule base already) then this implication is added to the rule base with a weight correcting the composed weight to the confidence of the implication in data.

In the next section 3, we discuss Łukasiewicz's composition operation and focus on situations when composed weights and confidences are strongly opposite. We solve this by adding to the rule set not only correcting rules, but also special turning rules.

Obtained results serve as a basis for modifications of the algorithm for the automatic rule set acquisition from categorical data using Łukasiewicz's composition function (section 4).

2 Method for rule set acquisition from data

We assume to have an observational categorical data set of an extent of m *objects* and n *attributes* (variables, features) at our disposal. These data are supposed to be in the form of a data matrix. Rows correspond to objects, columns correspond to attributes. Values of attributes (attribute - value pairs) describe different propositions (categories of objects) A_i.

Rule sets generated from data will be in our approach treated as sets of rules in the form of empirical implications $\alpha \Longrightarrow C$, where α is a conjunction of propositions $A_1 \wedge \ldots \wedge A_k$ and C is a fixed proposition – a given goal (class) C.

Let $F_D(\alpha)$ be the frequency of α in data D, i.e. the number of objects, which fulfil α in data D. For positive $F_D(\alpha)$, the empirical conditional probability $P_D(C|\alpha)$ of the implication $\alpha \Longrightarrow C$ in data D (called in data-mining terminology *confidence*) is:

$$P_D(C|\alpha) = \frac{F_D(\alpha \wedge C)}{F_D(\alpha)}. \tag{2}$$

The task of the rule set acquisition from the data consists in the question which of empirical implications existing in data are to be inserted into the resulting rule set, and with which weights. It depends not only on the data but also

- on the requirement of the accuracy with which the resulting rule set is to be in accordance with the data, and
- on the chosen inference mechanism which is to be used for processing the resulting rule set.

The accuracy of the rule set is controlled in the data D by a statistical test T of hypothesis that the conditional probability $P(C|\alpha)$ is equal to the weight inferred from the rule set using all rules satisfied by α.

The rule set is constructed by our original algorithm [9, 2] in a way analogous to the creating of an axiomatic theory. Here the state of axioms is given to the most simple statements so that all the other known statements of the domain could be inferred from them (the requirement of completeness). At the same time the redundancy is removed; statements derivable from other axioms are going not to be axioms (the requirement of independence).

The algorithm generates and tests empirical implications in the sequence according to the complexity or frequency of the left-hand side so the most reliable implications are generated first. This process starts with the "empty rule" with the weight equal to the relative frequency of the goal C in the data and stops after testing all existing implications. The algorithm generates every implication only once, and at the moment of testing some implication, all its sub-implications have been already processed.

During testing, the empirical conditional probability $P_D(C|\alpha)$ is computed. If it significantly differs from the composed weight (value obtained when composing weights of all sub-rules of the implication $\alpha \Longrightarrow C$) which have been inserted into the rule base already, then this implication is added to the rule set. Our first choice of a composition operation (see [9, 2, 1]) was the well-known Prospector pseudo-bayesian operation [3]:

$$x \oplus_P y = \frac{x * y}{x * y + (1 - x) * (1 - y)}$$

working on the unit interval $[0, 1]$. The operation should be used with respect to the syntactical dependencies among the rules by the application of the Möbius transform (according to the correction principle suggested by Hájek in [4]).

The original algorithm of the presented type was developed and implemented in the system ESOD (Expert System from Observational Data) [9], and nowadays is used in the system KEX (Knowledge EXplorer) with Prospector composition operation and usual statistical test. Several parameters can be used in the system to constrain the search space of implications, e.g. minimal required frequency of left-hand side. Satisfying results of testing this algorithm on several data sets were publish in [1]. Experiences with different data show that, typically, the acquired rule set consists of several percent of the number of all generated empirical implications and has a prediction ability better or comparable with other well known data-mining methods.

Let us underline that the described algorithm for automatic rule set acquisition from categorical data is general in the way that each composition function can serve as a basis for a modification of the algorithm. Now, we propose an application of the composition function derived from Łukasiewicz fuzzy logic with evaluated syntax $Ev_{\mathbf{Ł}}$ despite of its non-probabilistic character. It is motivated namely by the simplicity of

this composition function based on limited addition and subtraction of weights which is clearly understandable for experts.

3 Łukasiewicz's composition and turning rules

As a result of some theoretical considerations, an inference mechanism for uncertainty processing in rule-based systems based on complete Łukasiewicz's fuzzy logic with evaluated syntax Ev_{L} introduced by J. Pavelka (see [11], [12]) has been designed. It has been implemented previously in the System of Automatic Consultations (SAK), see [6, 10], nowadays in its follower New Expert System (NEST).

It uses several combination functions (for a general theory of combination functions see [4, 5]) which evaluate weights of formulas in the interval $[-1, +1]$. The composition $x \oplus_L y$ using Łukasiewicz's disjunction can be written as follows:

$$x \oplus_L y = \begin{cases} \min(1, x + y) & \text{for } x, y > 0 \\ \max(-1, x + y) & \text{for } x, y < 0 \\ x + y & \text{for } x \cdot y < 0 \end{cases} \tag{3}$$

An advantage of this operation is that the composition of weights is very simple: it is only the limited sum of weights. On the other hand, there are some problems arising when we try to apply this operation inside our method for automatic rule set acquisition from categorical data - namely the problem which is connected to situations when composed weights and confidences are strongly opposite. It means the difference between the required weight r based on confidence in data and the composed weight w is bigger than 1 so it is not possible to reach the weight r by any correcting weight r' from $[-1, +1]$. To solve this problem, we propose the idea of "turning rules".

First of all, let us describe the idea of turning rules on the simplest case: Let us consider a rule of the form

$$[A_1 \wedge A_2 \Longrightarrow C;\ r]$$

with the required weight r. Let $q = r_1 \oplus_L r_2$ be the weight composed by the Łukasiewicz's operator from weights of sub-rules

$$[A_1 \Longrightarrow C;\ r_1]$$

$$[A_2 \Longrightarrow C;\ r_2]$$

If q differs from r and $|r - q| < 1$ then the rule in question is inserted to the rule set with the corrected weight $r' = r - q$. As the result, the sequential composition of the weights r_1, r_2, r' gives the assumed weight r.

On the other hand, if $|r - q| > 1$ then it is not possible to reach the assumed weight r by any correcting weight r' from $[-1, +1]$. To solve this problem, the rule in question is inserted to the knowledge base two times: first as a *turning* rule with the weight $t = \mathrm{sgn}(r - q)$, and second as a *correcting* rule with the weight $r' = r - q - \mathrm{sgn}(r - q)$. As the result, the sequential composition of the weights r_1, r_2, t, r' gives the required weight r.

More general, the rule base will have in our case three sets of rules:
TR^+ is the set of turning rules with the weight 1,
TR^- is the set of turning rules with the weight -1,
CR is the set of correcting rules with weights from $[-1, +1]$.

The general composition function based on an application of Łukasiewicz's composition operation (3) will be defined in the interval $[-1, 1]$ by the function $GLOB^*$. It is an adoption of the function $GLOB$ (see [6, 10]) for working with both correcting and turning rules. The function $GLOB$ initially realizes composition of rules with weights $w_1, ..., w_n$ with the same conclusion C

$$\begin{aligned}&\mathrm{GLOB}([\alpha_1 \Longrightarrow C;\ w_1], ..., [\alpha_n \Longrightarrow C;\ w_n]) = \\ &\quad = [C;\quad \min(1, \textstyle\sum_{w_i>0} w_i) + \max(-1, \sum_{w_i<0} w_i)\ \ ]\end{aligned}$$

The function $GLOB^*$ is defined in such a way that $GLOB$ is applied sequentially according to lengths of rules. Let S_i be the subset of all rules of the length i. Let us define partial compositions:

$$t_i = GLOB(S_i \cap (TR^+ \cup TR^-)), c_i = GLOB(S_i \cap CR). \tag{4}$$

Then we apply the composition operation $x \oplus_L y$ (3) for the sequence of values $t_1, c_1, t_2, c_2, ...$ (from the left) to obtain the resulting composed weight.

4 Modified algorithm using Łukasiewicz's composition

The algorithm starts with the "empty rule" representing the relative frequency v_0 of C in the data D.

The "empty rule" is inserted to the rule set with the transformed weight $w_0 = 2v_0 - 1$.

The algorithm generates and tests empirical implications in the sequence according to the complexity or frequency of the left-hand side. Each implication $\alpha \Longrightarrow C$ is the candidate for a rule which is going to be included to the arising rule set.

During testing, the empirical conditional probability $v = P_D(C|\alpha)$ (confidence) is computed. Then the algorithm compares the candidate implication $\alpha \Longrightarrow C$ to existing rules. It composes (by the adopted function $GLOB^*$ in the interval $[-1, +1]$) weights of all sub-rules of the implication $\alpha \Longrightarrow C$ which have been inserted into the rule set already. For this, partial compositions $t_1, c_1, t_2, c_2, ...$ of the weights of turning, and correcting sub-rules, respectively, are calculated according their lengths and finally these weights are sequentially added by the composition operator (3) to the initial weight w_0. Let us denote the resulting composed weight by q.

If the empirical conditional probability $v = P_D(C|\alpha)$ of the candidate implication $\alpha \Longrightarrow C$ in the data D significantly differs from the estimation of its conditional probability obtained as the transformed value $\frac{q+1}{2}$ of the composed weight q (i.e. the hypothesis $P(C|\alpha) = \frac{q+1}{2}$ is rejected by the given statistical test T in the data D)

then this implication is added to the rule set. The computation of the correcting weight follows.

The empirical conditional probability v of the rule is transformed to the interval $[-1, +1]$ using formula $w = 2v - 1$. If $|w - q| < 1$ then the rule in question is inserted to the rule set with the corrected weight $w' = w - q$. As the result, the composition of the weights q, w' gives the weight w.

On the other hand, if $|w - q| > 1$ then the rule in question is inserted to the rule set two times: first as a turning rule with the weight $t = \mathrm{sgn}(w - q)$, and second as a correcting rule with the weight $w' = w - q - \mathrm{sgn}(w - q)$. As the result, the sequential composition of the weights q, t, w' gives the weight w.

Clearly, the modified algorithm has the following property which certifies that the resulting set of rules with the used composition function represents the data according the chosen statistical test:

Let a data matrix D, a goal C (some attribute - value pair or their combination), and a statistical test T be given. Let KB be the set of rules constructed by described algorithm. Let $A_1, \ldots, A_k$ be values of some different input attributes, and

$$[\alpha_1 \Longrightarrow C;\ w_1], ..., [\alpha_n \Longrightarrow C;\ w_n] \tag{5}$$

be all rules included in KB which are satisfied by $A_1, \ldots, A_k$. Let w^* be the composed weight

$$w^* = GLOB^*([\alpha_1 \Longrightarrow C;\ w_1], ..., [\alpha_n \Longrightarrow C;\ w_n]) \tag{6}$$

and $p^* = \frac{w^*+1}{2}$ be its transformation to the interval $[0, 1]$.

Then the hypothesis that the conditional probability $P(C|A_1 \wedge \ldots \wedge A_k)$ is equal to p^* is not rejected by the test T in the data D.

5 Conclusion

We discussed possible applications of Łukasiewicz's composition operation and the correction principle during our hierarchical construction of the rule base from data. More complex rules are added into the rule set only when the confidence of the rule in data significantly differs from the composed weight (value obtained when composing weights of all sub-rules of the implication which have been inserted into the rule base already). At this moment, a weight correcting the composed weight to the requested one is calculated. We proposed an addition of turning rules to the rule base in situations when composed and required weights are strongly opposite.

The resulting algorithm for automatic rule set acquisition from categorical data using Łukasiewicz's composition operation would be further elaborated and tested. We hope the presented approach can serve as a possibility to compose expert and data based knowledge in a common rule set using the same and simply understandable weights structure.

ACKNOWLEDGEMENT

This work was partially supported by the Ministry of Education, Youths and Sports of the Czech Republic (MSM 6138439910).

References

[1] Berka, P. (2012), Learning compositional decision rules using the KEX algorithm, *Intelligent Data Analysis 16*, 665–681.

[2] Berka, P. and Iránek, J. (1994), Automated knowledge acquisition for PROSPECTOR-like expert systems, In: *Machine Learning: ECML-94* (Bergadano, F. and Raedt, L.D., eds.), Springer Verlag, 339–342.

[3] Duda, R.O. and Gasching, J.E. (1979), Model design in the Prospector consultant system for mineral exploration, In: *Expert Systems in the Micro Electronic Age* (Michie, D., ed.), Edinburgh University Press.

[4] Hájek, P. (1985), Combining functions for certainty factors in consulting systems, *Int.J. Man-Machine Studies 22* (1985), 59–76.

[5] Hájek, P., Havránek, T. and Jiroušek, R. (1992), *Uncertain Information Processing in Expert Systems*, CRC Press, London.

[6] Iránek, J. (1991), Representation of expert knowledge as a fuzzy axiomatical theory, *International Journal of General Systems 20* (1991), 55–58.

[7] Iránek, J. (1994), Minimum knowledge base search from categorical data, In: *WUPES 1994* (Kramosil, I. and Jiroušek, R., eds.), Univ. of Economics, 77–86.

[8] Iránek, J. (2012), Some properties of evaluated implications used in knowledge-based systems and data-mining. *Journal of Systems Integration 3*, 17–23.

[9] Iránek, J. and Stejskal, B. (1988), Automatic acquisition of knowledge base from data without expert: ESOD (Expert System from Observational Data), In: *Proc. COMPSTAT'88*, Physica-Verlag, Heidelberg, 175–180.

[10] Iránek, J., Švenda, J. and Ferjenčík, J. (1989), Inference in expert systems based on complete multivalued logic, *Kybernetika 25*, 25–32.

[11] Novák, V., Perfilieva, I. and Močkoř, J. (1999), *Mathematical Principles of Fuzzy Logic*, Kluwer, Boston

[12] Pavelka, J. (1979), On Fuzzy logic I, II, III., *Zeischr. f. Math. Logik und Grundl. der Math. 25*, 45–52, 119–134, 447–464.

P-validity in a psychological context

Gernot D. Kleiter
Department of Psychology, University of Salzburg, Austria
gernot.kleiter@gmail.com

Abstract

The contributions throws a critical light on the application of Adams' p-validity in reasoning research. It investigates properties of p-valid inferences if the probabilities of the premises are point probabilities and not—as in the work of Adams and the related literature—interval probabilities with upper bounds equal one. Recently p-validity has been used as "a new standard" to evaluate human probability judgments. Judgments are classified as falling or not falling into "p-valid intervals" with upper bounds one. As in these experiments the probability assessments of the premises are point probabilities and not lower bounds of intervals with upper bounds one, this leads to classify incoherent and overconfident judgments as rational.

1 Introduction

For more than a millennium philosophers compared human reasoning with logical principles. In the last century psychologists developed new theories and methods but continued the comparison with the standards of classical logic. In the last decade, however, psychologists switched the perspective from classical logic to probability so that the old standards were not applicable any more. This included one of the most important standards of classical logic, the *validity* of inference rules:

> *If $\mathcal{A} = \{A_1, \ldots, A_n\}$ denotes a set of premises and B be a conclusion, then an inference rule is valid, $\mathcal{A} \models B$, iff it is impossible for all premises in $\mathcal{A}$ to be true and the conclusion B to be false.*

Looking for a similar standard that is applicable in the probabilistic approach psychologists hit on Adams' p-validity [1, 2, 3, 6, 5]. P-validity allows to classify probabilistic inference rules as "p-valid" or "p-nonvalid" analog to "valid" and "nonvalid" in classical logic. Adams introduced p-validity in probability logic as a surrogate for validity in classical logic. P-validity functions as a substitute, an "Ersatz", when "... 'probable' and 'improbable' are substituted for 'true' and 'false'." [3, p.1]
In addition, when p-validity is combined with the interpretation of the probability of conditionals as conditional probabilities, $P(\text{if } A \text{ then } B) = P(B|A)$, it has the nice property that it classifies the probabilistic versions of some classically valid but psychologically nonintuitive inference rules as p-nonvalid. The nonintuitive paradoxes

of the material implication, contraposition, or strengthening the antecedent are valid but p-nonvalid. Moreover, the set of p-valid inference rules [2, p.277, Definition 6] corresponds to the rules of system P [13, 17], a well known system of nonmonotonic logic. Would human reasoning be closer to such a system than to a system of classical logic [23]? P-validity has been discussed in the psychological literature, for example, in [21, 14, 23]. Recently it has been claimed to be a "new standard" to evaluate the rationality of probability judgments in human reasoning [25, 10]. The present contributions throws a critical light on the application of p-validity in reasoning research.

2 P-validity

Consider an inference with the premises $\mathcal{A} = \{A_1, \ldots, A_n\}$ and the conclusion B. Assume that interval probabilities $P(A_1) \in [\alpha'_1, 1], \ldots, P(A_n) \in [\alpha'_n, 1]$ are assessed for the premises. Let the lower bounds $(\alpha'_1, \ldots, \alpha'_n)$ of the assessment be coherent, that is, avoid sure loss. Following Adams, call the 1-complement of the probability of an event E its "uncertainty", $u(E) = 1 - P(E)$. Adams [5, p.38] introduced the following *uncertainty-sum criterion* to define probabilistically valid or "p-valid" inferences:

Definition 1 (P-validity) *"The uncertainty of the conclusion of a [probabilistically] valid inference cannot exceed the sum of the uncertainties of its premises" , that is,*

$$u(B) \leq \sum_{i=1}^{n} u(A_i) \quad \textit{or more explicitely} \quad u(B) \in \left[0, \min\left\{1, \sum_{i=1}^{n}(1-\alpha'_i)\right\}\right], \qquad (1)$$

where n denotes the number of premises and α'_i the lower probability of the i^{th} premise. A definition that is more explicit about the bounds is given in Adams' earlier papers [6, p.436] and in his 1975 book. Slightly reformulated it reads:

Definition 2 (Probability preservation) *The premise set $A_1, \ldots, A_n$ probabilistically entails the conclusion B with $P(B) \in [\gamma', 1]$ iff for all interval probabilities of the conclusion $[\gamma', 1]$ there exists a coherent interval assessment $[\alpha', 1]$ of the premises such that if $P(A_i) \in [\alpha', 1]$ for all A_i, then $P(B) \in [\gamma', 1]$.*

Often, when referring to the uncertainty-sum criterion, the literature is silent about the upper probability bounds of the premises. They are *implicitly* assumed to be all equal to 1. This may be obvious in a logical context. It led however to serious misunderstandings in the research on human reasoning. The uncertainty-sum criterion lures the understanding that the uncertainties are 1-complements of point probabilities while they actually are 1-complements of bounds of probability *intervals* [13]. As a consequence, what is called the "uncertainty" of the conclusions is actually the upper bound of an interval with lower bound equal to zero, $[0, u(B)'']$.
Definition 1 does *not generally* lead to *coherent* probabilities of the conclusion. Consider the following example:

Example 1 (Or-Introduction) From $P(A_1) \in [\alpha'_1, 1]$ *and* $P(A_2) \in [\alpha'_2, 1]$ *we infer the coherent probability interval* $P(A_1 \vee A_2) \in [\max\{\alpha'_1, \alpha'_2\}, 1]$ *which corresponds to the uncertainty interval* $u(B) \in [0, \min\{\alpha'_1, \alpha'_2\}]$. *The uncertainty-sum criterion, however, leads to the uncertainty interval* $[0, \min\{2 - (\alpha'_1 + \alpha'_2), 1\}]$, *the same bounds as for and-introduction.*

The uncertainty-sum criterion in Definition 1 is insensitive to the *logical form* of the conclusion. Adams [6] was well aware of this point and introduced an improved but less well-known definition. It involves the *essentialness* of the premises for the specific conclusion at hand. A premise A_i is essential if its removal from the set of premises makes the inference invalid. The degree of essentialness is denoted by $e(A_i)$ or by e_i for short.

Definition 3 (Probabilistic entailment) *The premise set* $\mathcal{A} = \{A_1, \ldots, A_n\}$ *probabilistically entails the conclusion* B *with* $P(B) \in [\gamma', 1]$ *iff for all interval probabilities of the conclusion* $[\gamma', 1]$ *there exists a coherent interval assessment* $[\alpha'_i, 1]$ *of the premises such that if* $P(A_i) = \alpha'_i \in [\alpha'_i, 1]$ *for all* A_i, *then* [1]

$$\gamma \in \left[\max \left\{ 0, 1 + \sum_{i=1}^{n} e_i \alpha_i - \sum_{i=1}^{n} e_i \right\}, 1 \right]. \tag{2}$$

If a premise A_i is not a member of any essential subset of $\mathcal{A}$ then its essentialness e_i is 0. Otherwise a premise belongs to one or more sets of essential premises. If the cardinality of the set with the smallest number of such premises to which A_i belongs is k, then $e_i = 1/k$.

The unweighted uncertainty-sum criteria in Definitions 1 and 2 are only coherent for inference forms in which each of the premises is essential with $e_i = 1$. This holds, e.g., for the MP, the MT, or the axioms of System P, but it does not hold generally. Essentialness has not been discussed in the psychological literature at all. But also in well-known philosophical sources [7, p. 131] it is not clear that the premises are interval probabilities and essentialness is not mentioned.

We have yet not given any justification for the p-validity formula. Why becomes the probability of the conjunction of the premises a surrogate of logical validity? Adams distinguishes formulas with and without conditionals. Conditional-free formulas are called "factual" formulas by Adams [3]. Probabilities in inference forms containing factual formulas only show a parallel to classical logic. In classical logic the *conjunction* of the premises of a valid argument implies its consequence:

$$\text{If } \{A_1, \ldots, A_n\} \models B, \text{ then } \bigwedge_{i=1}^{n} A_i \rightarrow B, \tag{3}$$

where $\models$ denotes entailment and $\rightarrow$ denotes material implication. The uncertainty-sum criterion with $e_i = 1$, $i = 1, \ldots, n$, is nothing else than the lower bound of the probability of a conjunction re-written in terms of 1-complements. Therefore

[1] For a proof see Theorem 3.5 in [22]. The Theorem omits the "max" and the 0, associated with the possibility of negative values.

all classically valid inference forms containing only factual formulas are p-valid by the uncertainty-sum criterion. If $e_i = 1$ for all A_i, $i = 1, \ldots, n$, the lower probability bound resulting from the probability of the conjunction of the premises is coherent. If not all $e_i = 1$ the weighting formula must be applied, otherwise the lower probability bounds are incoherent.

2.1 Premises with point probabilities

In every-day life interval assessments with upper probabilities 1 are unrealistic. Similarly, in a psychological experiment the participants would have to assess lower bounds and allow all upper bounds to be equal to 1. That is, each assessment would admit very high probabilities, including certainty. In the experiments of Singmann et al. [25] and of Evans et al. [10], for example, the participants did not assess intervals with upper probabilities equal to 1. The participants were asked to assess *point probabilities*, one judgment one number. The definition of p-validity becomes:

Definition 4 (Conjunction) *A logically valid inference form in which each premise A_i, $i = 1, \ldots, n$, is essential with $e_i = 1$ and which is conditional-free is p-valid iff the minimal probability of its conclusion, $\gamma' = P(B)$, is equal to or greater than the lower bound of the probability of the conjunction of its premises,*

$$\gamma' = P'(\bigwedge_{i=1}^{n} A_i) \geq \max \left\{ 0, \sum_{i=1}^{n} \alpha_i - (n-1) \right\} . \tag{4}$$

γ' is the coherent lower bound of the probability of the conjunction of a set of propositions when no assumptions about their dependence or independence are made.
What about upper bounds of the conclusions if the assessment of the premises is a point probability? Assuming $e_i = 1, i = 1, \ldots, n$ and conditional-free premises, the upper bound is obtained from the conjunction of the premises:

$$\gamma'' = P''(\bigwedge_{i=1}^{n} A_i) \leq \min\{\alpha_1, \ldots, \alpha_n\} . \tag{5}$$

Because of the *conjugacy* property

> *The upper bound of an interval probability is equivalent to 1 minus the lower bound of its complement, $P''(A) = 1 - P'(\neg A)$.*

p-validity may equally well be defined in terms of upper probabilities:

Definition 5 (Upper bounds) *A logically valid inference rule is p-valid iff the upper probability of its conclusion, $\gamma''(B)$, is less 1 minus the lower bound of the complement of its conclusion, that is, if $\gamma''(B) \leq 1 - \gamma'(\neg B)$.*

Because of the gap between γ'' and 1 p-validity not only protects against too low but also against too high probability judgments.

2.2 Premises containing conditionals

What if the premises contain conditionals and the probability of the conditionals are taken as conditional probabilities? To build the conjunction of the premises cannot be right here because the conditionals are "conditional events" of the form $E|H$ and conditional events are not ordinary propositions. The conjunction of conditional events does not follow the same principles as the conjunction of ordinary propositions.
The problem dissolves if we interpret the conditionals in the premises as *material implications*. Material implications belong to the realm of "ordinary" propositional calculus. What at first appears as an inconsistency—the conjunction of conditionals entails the conclusion—turns out to be the crucial point of Adams' introduction of conditional probabilities in probability logic. *In p-valid arguments conditional event interpretation leads to higher probabilities of the conclusions than material implications interpretation.*
If conditionals are interpreted as material implications, then de Finetti's Fundamental Theorem applies. Numerical solutions are found by linear or fractional programming [19, p.100 ff.], linear programming for conclusions without conditional events and fractional programming for conclusions with conditional events. Conditionals with zero probabilities of the conditioning event are important for improved algorithms in complex inferences. Several nested linear systems are analysed, each system corresponding to a "zero layer" [9].
To repeat, the p-validity bound is obtained if conditionals are interpreted as material implications and if the coherent lower probability of the conjunction of the premises is less than the lower bound for the conditional event interpretation.
Reasoning research has extensively studied the human interpretation of conditionals. A clear dominance of the conditional event interpretation was found in [11]. Conditionals are not interpreted as material implications. *P-validity may be seen as a relation between two interpretations of conditionals, material implication and conditional event.*
In the following sections we write $\beta_|$ or $\gamma_|$ to denote the probability of a conditional event and $\beta_\rightarrow$ or $\gamma_\rightarrow$ etc. to denote the probability of a material implication.

If $P(A) = \alpha_A$ and $P(B) = \alpha_B$, then the probability of material implication $P(A \rightarrow B) = \gamma_\rightarrow$ is in the interval

$$\gamma_\rightarrow \in [\max\{1 - \alpha_A, \alpha_B\}, \min\{1, 1 - \alpha_A + \alpha_B\}]. \tag{6}$$

Because $P(A \rightarrow B) = 1 - P(A \wedge \neg B)$ the minimum of $\gamma_\rightarrow$ is obtained if the probability of the conjunction of A and $\neg B$ is maximal, $P(A \wedge \neg B) = \min\{\alpha_A, 1 - \alpha_B\}$ so that its 1-complement is $\gamma'_\rightarrow = \max\{1 - \alpha_A, \alpha_B\}$. The maximum is obtained if $P(A \wedge \neg B)$ is minimized, i.e., if $P(A \wedge \neg B) = \max\{0, \alpha_A + (1 - \alpha_B) - 1\}$ so that the 1-complement is $\gamma''_\rightarrow = \min\{1, 1 - \alpha_A + \alpha_B\}$. If the conditional is interpreted as a conditional event we have:

If $P(A) = \alpha_A$ and $P(B) = \alpha_B$, then the probability of the conditional

event $P(B|A) = \gamma_|$ *is in the interval*

$$\gamma_| \in \left[\max\left\{0, \frac{\alpha_A + \alpha_B - 1}{\alpha_A}\right\}, \min\left\{1, \frac{\alpha_B}{\alpha_A}\right\}\right], \quad \text{if} \quad \alpha_A > 0 \qquad (7)$$

$$\text{and } \gamma_| \in [0,1] \text{ if } \alpha_A = 0.$$

For $0 < \alpha_A \leq 1$ the interval is obtained from $P(B|A) = P(A \wedge B)/P(A)$ and the bounds of the conjunction are $P(A \wedge B) \in [\max\{0, \alpha_A + \alpha_B - 1\}, \min\{\alpha_A, \alpha_B\}]$.
We illustrate the relationship between p-valid and p-nonvalid inferences on one hand and the interpretation of conditionals on the other hand by an elementary example.

Example 2 (Modus Ponens) From $P(A) = \alpha_A$, $P(A \to B) = \beta_\to$ *infer the p-validity bound* $\gamma'_\to = \max\{0, \alpha_A + \beta_\to - 1\}$. *The probabilities of the premises are required to be coherent, i.e.,* $\beta_\to \geq \alpha_A$. *We have the linear system*

$$\begin{aligned} x_1 + x_2 &= \alpha_A \\ x_1 + x_3 + x_4 &= \beta_\to, \quad \textit{and} \quad \sum_{i=1}^{4} x_i = 1, x_i \geq 0, i = 1, \ldots, 4\,, \end{aligned}$$

where $x_1 = P(A \wedge B)$, $x_2 = P(A \wedge \neg B)$, $x_3 = P(\neg A \wedge B)$, *and* $x_4 = P(\neg A \wedge \neg B)$; *The objective function is* $\gamma'_\to = x_1 + x_3$ *which obtains its minimum for* $x_3 = 0$, *so that* $\gamma'_\to = x_1 = P(A \wedge (A \to B))$, *the lower probability of the conjunction of the two premises, which is* $\max\{0, \alpha_A + \beta_\to - 1\}$. *For the conditional event interpretation we obtain* $\gamma_| \in [\alpha_A\beta_|, 1 - \alpha_A + \alpha_A\beta_|]$.

In the coherence approach it is completely "legal" to work with the zero probability $\alpha_A = 0$ leading to $\gamma_| \in [0,1]$. The solution—compare the Theorem of Total Probability—requires only the sum of two products, no ratios. In the relevant axiom of the coherence approach $P((A \wedge B)|H) = P(A|H)P(B|(A \wedge H))$ (see e.g. [9]) the conditional probabilities are primitive. In the Kolmogorov approach, however, the conditional probabilities are defined by the ratio of absolute probabilities. To take $P(B|A) = 1$ if $P(A) = 0$, as Adams does, leads to $P(B|A) + P(\neg B|A) = 2$. To say that $P(B|A)$ is undefined or illegal if $P(A) = 0$ implies to illegalize also $P(B|A)$ if $P(A) = 1$. Oaksford and Chater [21] assume that minor premises in MP, MT, DA, and AC, always have probability 1. However, if an MP with $P(A) = 1$ is fine, then a DA is undefined (and neither p-valid nor p-nonvalid) and vice versa.
Singmann et al. [25] and Evans et al. [10] employ the lower bound $\max\{0, \alpha_A + \beta_? - 1\}$ to evaluate the probability judgments of the participants in their experiments. The question mark indicates the unclear interpretation of the conditional. This lower bound is the coherent lower probability for the conjunction of the premises where conditionals are material conditionals, i.e., $\beta_?$ is a actually $\beta_\to$. In the experiments the upper probability used to evaluate the judgments is 1. The assessments of the premises were, however, not intervals with upper bound 1 but point probabilities. Therefore the upper bounds of the probability of the conclusions cannot be 1. If, however, in a psychological investigation p-validity "intervals" are determined it would be consistent

to determine the upper bounds with the material implication interpretation, that is, to work with $\gamma''_{\rightarrow}$ and not with $\gamma''_{?} = 1$. That way the two interpretations of conditionals might empirically be compared.
Adams' uncertainty-sum criterion does not apply to logically nonvalid inferences. In a psychological context, however, we may find an interpretation of the conditional which makes the inference rule valid. Let's take as an example DA. Here the *equivalence* interpretation leads to a valid argument form.

Example 3 (Denying the Antecedent) In classical logic DA is invalid, $\{A \rightarrow B, \neg A\} \not\models \neg B$. *The uncertainty-sum criterion leads to the lower probability* $\max\{0, \alpha_{\neg A} + \beta_{?} - 1\}$ *(? for the unknown interpretation). This formula is used by Singmann et al. [25] and Evans et al. [10]. But the result is different from both, from* $\gamma'_{\rightarrow}$ *and from* $\gamma'_{|}$. *As the DA is nonvalid the conjunction of its premises does not entail its conclusion. If the conditional* $A \rightarrow B$ *is interpreted as an equivalence, that is as* $A \leftrightarrow B$, *then the lower probability of the conclusion is equal to* $\max\{0, \alpha_{\neg A} + \beta_{\leftrightarrow} - 1\}$. *This biconditional makes the DA logically valid and p-valid.*

3 Correlated events

In a psychological context the judgment of correlation is often of similar importance as the judgment of probability. In an inference rule correlations may appear at two locations: at the premises and at the conclusion.
The 2×2 correlation ψ between the two binary events A and B is,

$$\psi = \frac{x_1 x_4 - x_2 x_3}{\sqrt{\alpha_A(1-\alpha_A)\alpha_B(1-\alpha_B)}}\,, \tag{8}$$

where $x_1 = P(A \wedge B)$, $x_2 = P(A \wedge \neg B)$, $x_3 = P(\neg A \wedge B)$, and $x_4 = P(\neg A \wedge \neg B)$; α_A and α_B constrain the value of ψ; lower and upper bounds are obtained from (8) and the four conjunction probabilities $x_1 \in [\max\{0, \alpha_A + \alpha_B - 1\}, \min\{\alpha_A, \alpha_B\}]$, $x_2 = \alpha_A - x_1$, $x_3 = \alpha_B - x_1$, and $x_4 = 1 - (x_1 + x_2 + x_3)$.

3.1 Correlation in the conclusion

The probability of a single *if A then B* sentence carries little or no information about the correlation between the two events. From the material implication $P(A \rightarrow B) = \beta_{\rightarrow}$ we infer $\psi \in [-1, \frac{\beta_{\rightarrow}/2}{1-\beta_{\rightarrow}/2}]$ and from the conditional event $P(B|A) = \beta_{|}$ we can only infer the vacuous interval $\psi \in [0, 1]$. This changes of course when premises are added.
From the premises of the rules we infer the lower and upper correlations for the material implication interpretation, $[\psi'_{\rightarrow}, \psi''_{\rightarrow}]$, and for the conditional event interpretation, $[\psi'_{|}, \psi''_{|}]$, of the conditional in the premises. The results are obtained with the help of the lower and upper probabilities of the conclusions, i.e., $P(B), P(\neg A), P(\neg B)$, and $P(A)$ for MP, MT, DA, and AC, respectively. They allow to determine x_1, x_2, x_3, and x_4 which are required to determine ψ by Formula 8.

The Figures 1 and 2 show two numerical examples. The probabilities of the minor premises $P(A)$, $P(\neg B)$, $P(\neg A)$, and $P(B)$ are fixed at $\alpha = 0.5$. The probability of the major premise, $\beta = P($ if A then $B)$, is represented along the X-axis. The continuous lines show the results for the conditional event interpretation, the dashed lines those for the material conditional.
For MP and DA the results for the material implication interpretation are identical because of the symmetry of α and $1 - \alpha$ around .5. For all four inferences the upper correlation increases from $\psi_{\rightarrow} = 0$ approximately linearly up to $\psi_{\rightarrow} = 1$. At $P(A \rightarrow B) = .5$ the correlation must be negative that is, the upper bound, $\psi''_{\rightarrow} \leq 0$. The least informative rule is the AC with the widest intervals for $\psi_{\rightarrow}$. For the conditional event interpretation the MP, the MT and the AC infer positive correlations with $\psi'_{|} \geq 0$. The DA is very noninformative.
Inferences about the correlation allow to state qualitative properties like "always positive", "always negative", and "noninformative". The conditional event interpretation is more intuitive than the material implication interpretation. For the MP and the AC the conditional event interpretation uniformly dominates the material implication interpretation, $\psi'_{|} \geq \psi'_{\rightarrow}$ and $\psi''_{|} \geq \psi''_{\rightarrow}$.
The distinction between p-valid and p-nonvalid inferences rules does not identify strong or weak conclusions about correlations.

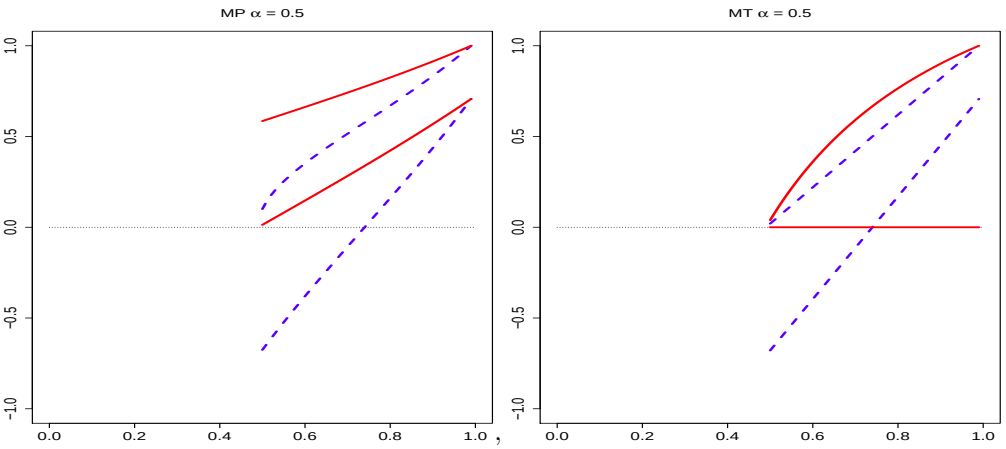


Figure 1: Left: MP, ψ on the Y-axis, $\alpha = P(A) = .5$ and $\beta = P($ if A then $B)$ on the X-axis. Right: MT, $\alpha = P(\neg B) = .5$. Line: Material implication. Dashed: Conditional event interpretation.

3.2 Correlation in the premises

Experiments on human reasoning often investigate inferences with *content-lean* material, like "If there is an A on one side of the blackboard, then there is a B on the other side.". Such a conditional does not carry information about the dependence or independence of the involved events. If however content-rich material is presented,

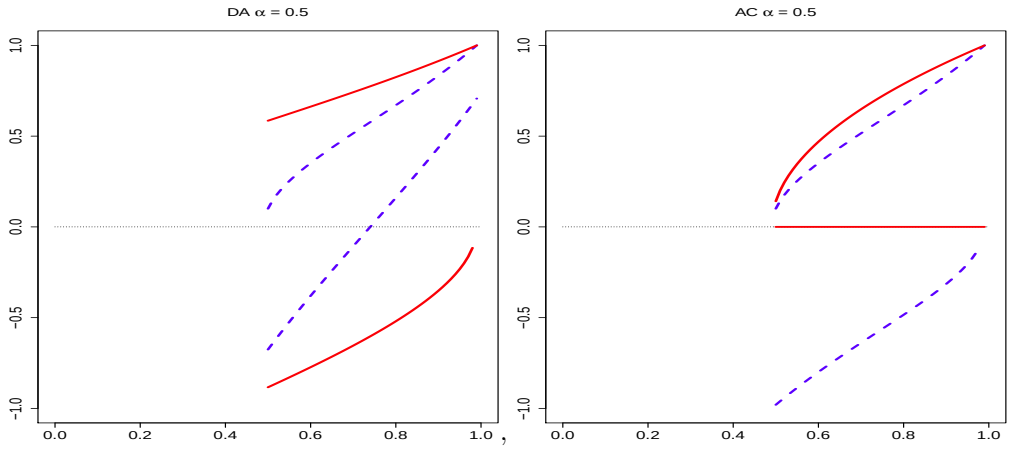


Figure 2: Left: DA, ψ on the Y-axis, $\alpha = P(\neg A) = .5$ and $\beta = P($ if A then $B)$ on the X-axis. Right: AC, $\alpha = P(B) = .5$. Line: Material implication. Dashed: Conditional event interpretation.

then the participants have background knowledge that will enter the inference process. Especially *if-then* sentences in the premises will activate beliefs about causal and correlational dependencies. Similarly, in every-day arguments events are usually supposed to be correlated.

A first anchor is independence. It is plausible to suppose that in every-day conversation we understand "if A then B" as "$P(B|A) > P(B)$". If in a psychological experiment the participants assess point probabilities $P(A)$ and $P(B|A)$ in an MP and the participants believe that A and B are independent, then their conclusion should be $P(B) = P(B|A)$. More generally, With some backward-engineering it is easy to infer an underlying correlation. For the MP with conditional event interpretation we apply formula 8 with

$$x_1 = \alpha\beta_| \quad x_2 = \alpha(1 - \beta_|) \tag{9}$$

$$x_3 = \gamma - x_1 \quad x_4 = 1 - \gamma - x_2 \tag{10}$$

The same may be done for MT, DA, and AC. In all these cases the judgment of point probabilities reveals the perceived correlation between A and B.

4 Distributions instead of intervals

The interval assessment of the premise probabilities raises the question of coherence. Usually not all combinations of point probabilities, each one belonging to its according interval, are jointly coherent. To handle coherence for interval probabilities Gilio introduced the concept of *generalized coherence* (g-coherence) [12], [8, Definition 2].

Definition 6 (g-coherence) *The vector of lower bounds* $(\alpha'_i, \ldots, \alpha'_n)$ *is g-coherent*

iff there exists a precise coherent assessment $(\alpha_i, \ldots, \alpha_n)$ *such that* $(\alpha_i \in [\alpha_i', \alpha_i''], \ldots, \alpha_n \in [\alpha_n', \alpha_n''])$.

As conjugacy holds, $P(E)'' = 1 - P(\neg E)'$, the upper bounds need not be included in the definition.
The lower and upper bounds of the premise probabilities α_i span an n-dimensional hypercube with the volume $\prod_{i=1}^{n}(\alpha_i'' - \alpha_i')$. Coherence defines a convex polyeder within the hypercube. Each point in the polyeder corresponds to a vector of a coherent point probability. G-coherence requires the polyeder not to be empty. The ratio of the volume of the polyeder and the hypercube is a measure of the degree of coherence for a given pair of vectors of lower and upper probabilities.
If we treat the α_i as random variables, introduce rectangular density functions on the $[\alpha_i', \alpha_i'']$ intervals, $f(\alpha_i) = 1/(\alpha_i'' - \alpha_i')$, and if we assume that the α_i are stochastically independent, then volumes in the hypercube correspond to a probability measure. The volume of the coherent subspace measures the (second order) probability of being coherent.
It is however more general to replace the rectangular by more flexible distributions, to replace the intervals $[\alpha_i', \alpha_i'']$ by the full range of the unit interval $[0, 1]$, and to replace the independence assumption by an appropriate measure of probabilistic dependence [15]. The resulting structure is a *vine* structure [18]. It is characterized as follows:

1. The imprecise uncertainty of the n premises is modeled by a multivariate probability density on the simplex $[0, 1]^n$.

2. The (marginal) uncertainty of each premise is described by an appropriate probability density, e.g., a beta distribution.

3. The pairwise (unconditional and conditional) stochastic dependencies are characterized by copulas. *Regular vines* allow a pairwise decomposition of the joint distributions.

4. Practical numerical analyses are performed by stochastic simulation.

The architecture corresponds to a *stochastic response model*. An individual represents his or her uncertainty by a distribution and when asked for a probability judgment responds with a random number generated by the distribution [15].

5 Discussion

All classically valid inference forms of propositional calculus which are conditional-free are p-valid. Of those containing conditionals a subset is p-nonvalid, most typically the paradoxes of the material implication, but also strengthening the antecedent, transitivity, contraposition, or-to-if (from "$A \vee B$" infer "if $\neg A$ then 'B"). Psychologically the p-nonvalid rules are just those which are nonintuitive (except for transitivity!). The p-valid rules correspond to the rules of System P [17, 13, 23], a prominent system of nonmonotonic logic and attractive for modeling human reasoning. Moreover, the two kinds of interpretation of conditionals—conditional event and material

implication— are the decisive criterion to distinguish p-valid from p-nonvalid rules. The nonintuitive rules are filtered out by the stronger conditional event interpretation. This is the reason why Adams used conditional probabilities for the probability of conditionals. Both, p-validity and conditional probabilities, go hand in hand.

Recent psychological studies [10, 25] used p-validity to evaluate human judgments as falling into "p-valid" intervals or not. The intervals are claimed to be a new standard of rationality. These studies do not see that Adams assigns interval probabilities with upper probabilities equal to 1 to the premises, that is, not point probabilities as in the judgments of the participants in the experiments. For inference rules like the MP or the MT, where all premises have degrees of essentialness equal to 1, Adams' uncertainty-sum is identical to the lower probability of the conjunction of the premises if the conditionals are interpreted as material conditionals. If the inferences are not content-lean but involve causal knowledge, background knowledge about correlations narrows down the intervals of the probability of the conclusion, leading to point probabilities in the case of precise correlations. Psychologically it is highly plausible to abandon both, models with point probabilities and models with interval probabilities, and to replace them with models in which imprecision is represented by continuous probability density functions. The strict classification as "coherent" and "non-coherent" dissolves and is replaced by distributions on degrees of coherence.

The coherence approach has an elegant method to establish the bridge between classical logic and probability. It does necessarily start from a Boolean algebra. If the premises are *logically* dependent this is directly taken into account by removing impossible constituents, those that are forbidden by the logical dependence right at the beginning of any analysis [9].

Adams distinguishes different kinds of probability preservation, among them certainty preservation [4]. "A is a *strict [certainty preserving] consequence* of S ... if and only if for all probability functions P ... if $P(B) = 1$ for all B in S, then $P(A) = 1$." [2, p. 274] McGee [20] observes that this criterion falls back to material implication: "The strictly valid inferences are not those described by Adams' theory, but those described by the orthodox theory, which treats the English conditional as the material conditional. This raises an ugly suspicion. The failures of the classical valid modes of inference appear only when we are reasoning from premises that are less than certain ... to a conclusion that is also less than certain." [20, p.189] This is a consequence of Adams' conception of conditional probability as defined by $P($ if A then $B)$ as $P(A \wedge B)/P(B)$ if $P(B) \neq 0$ and as 1 if $P(B) = 0$, i.e., he "...assigns the conditional the probability 1 when the conditional probability is undefined" [20, p. 190]. McGee proposed to employ Popper functions, but zero probabilities are directly addressed in the coherence approach.

P-nonvalid rules may lead to informative and important probabilities of the conclusion— if the judgments are coherent. The distinction between p-valid and p-nonvalid rules does not touch the entailment relation based on coherence. In probabilistic inference coherence is the gold standard. In models of human reasoning p-validity is a relict of classical logic.

References

[1] E. W. Adams. The logic of conditionals. *Inquiry: An Interdisciplinary Journal of Philosophy*, 8:166–197, 1965.

[2] E. W. Adams. Probability and the logic of conditionals. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 265–316. North-Holland, Amsterdam, 1966.

[3] E. W. Adams. *The Logic of Conditionals*. Reidel, Dordrecht, 1975.

[4] E. W. Adams. Four probability-preserving properties of inferences. *Journal of Philosophical Logic*, 25:1–24, 1996.

[5] E. W. Adams. *A Primer of Probability Logic*. CSLI Publications, Stanford, 1998.

[6] E. W. Adams and H. Levine. On the uncertainties transmitted from premises to conclusions in deductive inferences. *Synthese*, 30:429–460, 1975.

[7] J. Bennett. *A philosophical guide to conditionals*. Oxford University Press, Oxford, 2003.

[8] V. Biazzo and A. Gilio. A generalization of the Fundamental Theorem of de Finetti for imprecise conditional probability assessments. In *1st International Symposium on Imprecise Probabilities and Their Applications*. Electronic Version at http://decsai.ugr.es/ smc/isipta99/proc/009.html, Ghent, Belgium, 29 June–2 July 1999.

[9] G. Coletti and R. Scozzafava. *Probabilistic logic in a coherent setting*. Kluwer, Dordrecht, 2002.

[10] St B. T. Evans, V. Thompson, and Over D. E. Uncertain deduction and conditional reasoning. *Frontiers in Psychology. Cognition*, 6:398, 2015.

[11] A. J. B. Fugard, N. Pfeifer, B. Mayerhofer, and G. D. Kleiter. How people interpret conditionals: Shifts toward the conditional event. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37:635–648, 2011.

[12] A. Gilio. Probabilistic consistency of conditional probability bounds. In B. Bouchon-Neunier, R. R. Yager, and I. A. Zadah, editors, *Lecture Notes in Computer Science 945*, pages 200–209. Springer, Berlin, 1995.

[13] A. Gilio. Probabilistic reasoning under coherence in system P. *Annals of Mathematics and Artificial Intelligence*, 34:5–34, 2002.

[14] P. N. Johnson-Laird, S. S. Khemlani, and G. P. Goodwin. Logic, probability, and human reasoninig. *Trends in Cognitive Sciences*, xx:1–14, 2015.

[15] G. D. Kleiter. Modeling biased information seeking with second order probability distributions. *Kybernetika*, in print, 2015.

[16] G. D. Kleiter, M. E. Doherty, and G. L. Brake. The psychophysics metaphor in calibration research. In P. Sedlmeier and T. Betsch, editors, *Etc. Frequency Processing and Cognition*, pages 239–255. Oxford University Press, Oxford, UK, 2002.

[17] K. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[18] D. Kurowicka and R. Joe. *Dependence Modeling: Vine Copula Handbook*. World Scientific, Singapure, 2011.

[19] F. Lad. *Operational Subjective Statistical Methods*. Wiley, New York, 1996.

[20] V. McGee. Learning the impossible. In E. Eells and B. Skyrms, editors, *Probability and Conditionals*, pages 179–199. Cambridge University Press, Cambridge, UK, 1994.

[21] M. Oaksford and N. Chater. *Bayesian Rationality. The Probabilistic Approach to Human Reasoning*. Oxford University Press, Oxford, 2007.

[22] J. B. Paris and A. Vencovská. *Pure Inductive Logic*. Cambridge University Press, Cambridge, UK, 2015.

[23] N. Pfeifer and G. D. Kleiter. Coherence and nonmonotonicity in human reasoning. *Synthese*, 146:93–109, 2005.

[24] N. Pfeifer and G. D. Kleiter. Uncertain deductive reasoning. In K. Manktelow, D. E. Over, and Elqayam S., editors, *The science of reasoning: A Festschrift for Jonathan St B.T. Evans*, pages 145–166. Psychology Press, Hove, UK, 2010.

[25] H. Singmann, K. C. Klauer, and D. E. Over. New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology*, 5(PMC4029011):316, 2014.

Diagnostic problem without marginals

Otakar Kříž
Prague, Czech Republic
e-mail: o.kriz@upcmail.cz

Abstract

An algorithm (Symptom Proximity) is suggested for solving discrete **diagnostic problem**. It is based on probabilistic approach to decision-making under uncertainty, however, it does not use knowledge integration from marginal distributions.

1 Introduction

1.1 The layout of the paper

1. There are historical reminiscences explaining the position of the suggested method in a broader context in subsection 1.2.

2. Basic notions are defined including the formulation of the *diagnostic problem* and describing the role of the *statistical file* $\mathcal{F}$ in subsection 1.3.

3. The essential features of the algorithm SP are laid down in section 2.

4. SP is described in a symbolic programming language in section 3.

5. On the basis of this description, computational complexity C_{SP} of SP in terms of "length" $l = |\mathcal{F}|$ of the file $\mathcal{F}$ and of its "width" $n = |\{\xi_1, \xi_2, \cdots \xi_n\}|$ is estimated and verified experimentally for different values of l and w in section 4.

6. "Discernment power" of SP (i.e. absolute values or percentage of wrong classifications) is tested for different "apertures" (sets of symptom variables whose values are disclosed to SP as evidence). Testing is performed both via method "leave one out" as well as on all data. The results are compared with a simple marginal-based algorithm under the same testing conditions in section 5.

7. Features of SP sorted as "pros" and "cons" are summarized in section 6 .

1.2 Historical background

Firat attempts for machine-assisted decision-making under uncertainty are marked by rule-based expert systems Mycin and Prospector in early eighties. Weights in rules were interpreted as conditional probabilities. But the way the rules were combined was not probabilistic ones. The same held for systems with fuzzy number approach. At that time, Albert Perez in [9]) raised the requirement that partial knowledge should be "integrated" intensionally i.e. using the concept of theoretical joint distribution P. Knowledge was understood as probability or conditional probability elicited either from experts or observed from experiments. The best way to keep it, at least partially, complete and homogenous was to assume that it comes in form of less-dimensional distributions that were supposed to be marginals of the theoretical joint P. Thanks to smaller sizes, marginals could be estimated from available data. The main effort in the subsequent research was concentrated on the way how to assemble effective approximations of the joint P. The formulation of the task was known as *marginal problem* already in [4] and its specific solution was suggested even before in [2]. Different models, connected with names like Lauritzen, Spiegelhalter, Dempster, Shafer, Pearl, Dawid, were studied with assumptions about conditional independence of variables appearing in P that helped to integrate the marginals. At present, there exist professional software packages (e.g. Hugin) supporting the decision-making on commercial basis. As, beside different algorithms, even the selection of proper marginals may be a problem of its own (see e.g. [7] and [8]), this paper tries to study an alternative to the marginal approach.

1.3 Basic notions

Let $(\Omega, \mathcal{X}, P)$ be a probabilistic space,
$\boldsymbol{\eta} = \boldsymbol{\xi_0}, \boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \ldots \boldsymbol{\xi_n}$ be finite sets and
$\xi_r : (\Omega, \mathcal{X}, P) \longrightarrow (\boldsymbol{\xi}_r, 2^{\boldsymbol{\xi}_r})$ for $r = 0, 1, 2, \cdots n$ be measurable functions.

Though the topic is defined in a formal way, the names of objects in the universe of discussion (e.g. diagnosis, symptoms etc.) are taken from the field of medicine to give them a semantical interpretation and ease up understanding of basic notions and character of their interaction.

The mutual behaviour of all random variables $\eta, \xi_1, \xi_2 \cdots \xi_n$ is described by a theoretical joint probability distribution $P_{\eta\,\xi_1\xi_2\cdots\xi_n}$.

Decision making under uncertainty with probabilistic background can be interpreted as the diagnostic problem with the following formulation:

Diagnostic problem: Find the diagnosis $d(s_1, s_2 \cdots s_n) \in \boldsymbol{\eta}$ that is the most probable (according to the $P_{\eta\xi_1,\xi_2\cdots\xi_n}$) on the set

$$\{\omega \in \Omega \mid \xi_1(\omega) = s_1 \;\&\, \xi_2(\omega) = s_2 \;\&\; \cdots \xi_n(\omega) = s_n\}$$

for a given (i.e. observed) arbitrary combination $(s_1, s_2 \cdots s_n)$ of values of *symptom variables* from the cartesian product $\boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$.

If we wish to predict the values of diagnostic variable η, the conditional probability $P_{\eta|\xi_1\xi_2\ldots\xi_n}$ (derivable from $P_{\eta\,\xi_1\xi_2\ldots\xi_n}$) should be used instead of $P_{\eta\,\xi_1\xi_2\ldots\xi_n}$.

Optimal decision: The value of diagnosis d from $\boldsymbol{\eta}$ that should be selected if the values of symptom variables are $(s_1, s_2 \cdots s_n)$ to keep the wrong classification of d as low as possible), called Bayes solution, is for each $(s_1, s_2 \cdots s_n) \in \boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$ given by the formula

$$d_{opt}(s_1, s_2 \cdots s_n) = \underset{d \in \boldsymbol{\eta}}{\operatorname{argmax}}\ P_{\eta\,|\xi_1\xi_2\ldots\xi_n}(d|s_1, s_2 \cdots s_n) \tag{1}$$

So far the theory. 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.

Marginal problem: Using the concept of *marginal problem*, see [4], *knowledge base* $\mathcal{K}$ is given as a set of "low-dimensional" distributions (e.g. number of variables in the distribution does not exceed e.g. 10.), postulated as theoretical *marginal distributions* of the $P_{\eta\xi_1,\xi_2\ldots\xi_n}$. Beside the marginals, there are usually made assumptions about conditional independence holding between groups of random variables. It is interesting that the topic was so attractive that it was addressed in several waves, usually after 20 years. Original and interesting ideas were not just the product of the last two decades but go back much deeper. See e.g. [2], [4],[3], [1]. Instead of the unknown $P_{\eta\xi_1\xi_2\cdots\xi_n}$, we try to construct (from the marginals) its suitable approximation $\hat{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ that could play its role in the formula (1).

If existence of marginals is postulated, it is natural to ask where do they come from. Therefore, another notion should be specified.

Statistical file F: Let $(\omega_1, \omega_2, \cdots \omega_s)$ be a sequence, where individual $\omega_i \in \Omega$ denote realizations of a random selection from Ω, then the sequence

$$(\eta(\omega_l), \xi_1(\omega_l), \xi_2(\omega_l) \cdots \xi_n(\omega_l))_{l=1}^{s}$$

of points in cartesian product $\boldsymbol{\eta} \times \boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$ is a *statistical file* F of size s (i.e. $s = |\mathcal{F}|$) and $(\mathcal{F})_r$ is the r-th member of $\mathcal{F}$.

There exists a taciturn assumption that decision making about a concrete case (patient) should be very fast (about 1 sec/pers.). On the other hand, longer time (e.g. hours of CPU time) devoted to selecting and populating the marginals (in the learning phase) is tolerable. This may be one of the reasons why "marginal approach" is the standard way.

However, using marginals for "integrating" $\hat{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ and its subsequent conditioning need not be mandatory for solving the diagnostic problem.

2 Basic idea of SP algorithm

An algorithm, called SP (Symptom proximity), tries to construct necessary conditional probabilities directly from available statistical data file $\mathcal{F}$. Basic idea of SP

can be explained by an assumption "Patients with similar smptoms should have a similar diagnosis". Hence, the name of the algorithms SP interpretes the similarity as a proximity in the sense of a very natural metrics.

Proximity metrics ρ :

$$\rho : \boldsymbol{\Xi} \times \boldsymbol{\Xi} \longrightarrow \boldsymbol{R} \qquad (\mathbf{u}, \mathbf{v}) \longmapsto n - \sum_{i=1}^{n} \delta(\,(\mathbf{u})_i, (\mathbf{v})_i)$$

where $\delta(\cdot,\cdot)$ is the Kronecker function and $(\mathbf{u})_i$ is the i-th component of the sequence $\mathbf{u}$

The mapping ρ is a metrics (i.e. reflexivity, symmetry, triangular inequality) on $\boldsymbol{\Xi}$ that can be used for defining equivalence classes on $\boldsymbol{\Xi}$. For each $\mathbf{v} \in \boldsymbol{\Xi}$, there exist $n+1$ sets $C_0(\mathbf{v}), C_1(\mathbf{v}), \cdots C_{n+1}(\mathbf{v})$, where $C_k(\mathbf{v}) = \{\mathbf{u} \in \boldsymbol{\Xi} \mid \rho(\mathbf{u}, \mathbf{v}) = k\}$.

The next step is to estimate $P(C_k(\mathbf{v}))$. It can be done, in a natural way, using available data (i.e. the statistical file $\mathcal{F}$).

$$P\left(C_k(\mathbf{v})\right) = \sum_{j=1}^{|\mathcal{F}|} \delta(\rho(((\mathcal{F})_j)_{\boldsymbol{\Xi}}, \mathbf{v}), k)/|\mathcal{F}|$$

where $(\mathcal{F})_j$ is the j-th vector from file $\mathcal{F}$ i.e. $(\mathcal{F})_j \in \boldsymbol{\eta} \times \boldsymbol{\Xi}$. Similarly, $((\mathcal{F})_j)_{\boldsymbol{\Xi}}$ is that part of the j-th vector $(\mathcal{F})_j$ that corresponds to symptom variables i.e. $((\mathcal{F})_j)_{\Xi} \in \boldsymbol{\Xi}$. We are interested in the set $C_k(\mathbf{v})$ with smallest k but at the same time such that $P\left(C_k(\mathbf{v})\right) > 0$. Let us denote this optimal k as k_0.

Finally, the conditional probability of η can be defined on $C_{k_0}(\mathbf{v})$

$$P\left(\eta | C_{k_0}(\mathbf{v})\right)(d|\mathbf{v}) = 1/(|\mathcal{F}| \cdot P(C_{k_0}(\mathbf{v})) \sum_{j=1}^{|\mathcal{F}|} \delta(\rho(((\mathcal{F})_j)_{\boldsymbol{\Xi}}, \mathbf{v}), k_0) \cdot \delta(((\mathcal{F})_j)_{\eta}, d)$$

If $\mathbf{v} = (s_1, s_2, \cdots s_n) \in \boldsymbol{\Xi}$, we may approximate the conditional probability $P_{\eta\,|\xi_1\xi_2\ldots\xi_n}(d|s_1, s_2 \cdots s_n)$ appearing in formula(1) by the $P\left(\eta|C_{k_0}(\mathbf{v})\right)(d|\mathbf{v})$ so that $P_{\eta\,|\xi_1\xi_2\ldots\xi_n}(d|s_1, s_2 \cdots s_n) = P\left(\eta|C_{k_0}(\mathbf{v})\right)(d|\mathbf{v})$ and formula (1) can be applied as the decision rule in the SP algorithm.

The algorithm SP is presented in a symbolic programming language in section 3. The complexity of the algorithm SP will be defined, in section 4, as a function of size $|\mathcal{F}|$ of the data file $\mathcal{F}$ and as a function of number n of symptom variables. The complexity is verified on real data by measuring time required for making decision for one person.

The decision quality (or discernment power) is dealt with in section 5. In principle, it is the number of wrong classification what is measured. However, it may defined more formally:

Let $\mathcal{L} \subset \mathcal{F}, f \in \mathcal{F}, \mathbf{v} \in \boldsymbol{\Xi}$. Further, let $\mathrm{SP}(\mathcal{L}, \mathbf{v}) \in \boldsymbol{\eta}$ denote decision of SP when evidence (about a patient) is $\mathbf{v}$ and algorithm SP has the "learning" file at his

disposal. Then, "discernment power" of SP can be measured by percentage of wrong classification either as

$$100\left[1-1/|\mathcal{F}|\sum_{j=1}^{|\mathcal{F}|}\delta(\mathrm{SP}(\mathcal{F},((\mathcal{F})_j)_\Xi),((\mathcal{F})_j)_\eta)\right]$$

or with the formula

$$100\left[1-1/|\mathcal{F}|\sum_{j=1}^{|\mathcal{F}|}\delta(\mathrm{SP}(\mathcal{F}\backslash(\mathcal{F})_j,((\mathcal{F})_j)_\Xi),((\mathcal{F})_j)_\eta)\right]$$

.

This second approach is referred to as "Leave one out" technique.

The results will be compared with one simple algorithm using the "marginal approach" in section 5.

3 Description of SP in a symbolic language

SP algorithm can be used in different roles. It may be a simple "one-shot" decision-making, repeated decision-making for different apertures, using SP in a general testing scheme or it may be required for specific testing via "Leave one out" technique.

Instead of using one highly parameterized form of SP algorithm, it seems better, from didactical reasons, to use several stand-alone modifications. However, only the most simple version, under the name *function SP*, will be presented in this paper. Specific modifications built on its basis (and entitled SPL and SPA) will be mentioned in other sections.

The following symbolic description is kept as simple as possible.

First, though the variables have their specific denotation reflecting their semantics, they are coded as integers or arrays of integers to make SP faster.

Second, tests and resulting exceptions in inconsistent situations such as $|\mathcal{L}| = 0$ or $|\boldsymbol{\eta}| = 0$ are omitted! Function **SP** returns the value $d_{opt}(t)$ for each $t = (s_1, s_2 \cdots s_n) \in \boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$

```
 1    function SP (t)
 2      read L  ⟶  L(0 − n, 1 − |L|)
 3       for j = 1, |η|
 4            for i = 1, n
 5              LD(i, j) = 0
 6            next i
 7       next j
 8       maxcount = 0
 9       for l = 1, |L|
10         count = 0
11            for j = 1, n
12              if L(j, l) = t(j) then
```

```
13                  count = count + 1
14                endif
15              next j
16              if maxcount < count then
17                maxcount = count
18              endif
19              d = L(0, l)
20              LD(count, d) = LD(count, d) + 1
21          next l
22          max = 0; d_opt = 0
23          for j = 1, |η|
24              val = LD(maxcount, j)
25              if max < val then
26                max = val
27                d_opt = j
28              endif
29          next j
30          SP = d_opt
31      end
```

Comments to the code of SP:

l.1 expresses that SP is a function $SP : \Xi \longrightarrow \boldsymbol{\eta}$ i.e. accepts as argument the vector t and returns the optimal diagnosis d_{opt}.

l.2 learning file $\mathcal{L}$ is stored in an array L. The value "0" in first dimension is for values of η.

l.3 - l.7 sets zero values to the array LD (level distance) where metrics will be stored in the sequel.

l.9 - l.21 For each $l \in \mathcal{L}$, number of symptom variables with coinciding values (symptoms) is calculated in the variable *count*. Increasing LD($count, \eta$(l)) by one increases chances of diagnosis η(l) to become optimal d_{opt} if the decision should take place at the level *count*.

l.16 - l.18 stores in *maxcount* the up-to-now achieved maximal number of coincidences.

l.23 - l.30 finds in LD(maxcount,j) such diagnosis d_j that would, on the level *maxcount*, define the winning dignosis d_{opt}. Naturally, if the number of cases from $\mathcal{L}$ is small (and that would result in objections from statistical point of view), it is possible to perform search for optimal diagnosis on a level *count* smaller than *maxcount* that would have more objects than level *maxcount*. Or even, it is possible to sum $LD(ct, j)$ for $ct = count$ to *maxcount* in an array D(1-$|\boldsymbol{\eta}|$) and search for d_{opt} in this array. (However, this modification of SP is not available in the presented version.)

The link of the code with previous formal description may be made more clear if we realize that the value in $LD(maxcount, j)$ is proportional to the probability $P\left(C_{maxcount}(\mathbf{t})\right)$ for the diagnosis j. Variable *maxcount* corresponds to k_0.

4 Computational complexity

It should be mentioned that experimenting with algorithms was performed on a statistical file $\mathcal{F}$, from the field of rheumatology with 1089 patients. Diagnosis variable η contained 4 diagnosis and there were 34 symptom variables whose ranges had cardinalities from 2 to 9. That way, no generation of artificial examples was necessary. Nevertheless, this choice has no influence on the substance of SP.

Complexity C_{SP} of SP algorithm can be mesured with respect to the number of symptom variables n, number $|\mathcal{L}|$ of objects in the learning file $\mathcal{L}$ and with respect to the number $|\boldsymbol{\eta}|$ of diagnoses i.e. $C_{SP} = C(n, |\mathcal{L}|, |\boldsymbol{\eta}|)$. Due to the simple structure of SP, C_{SP} can be estimated directly:

l.2	$c_1 * n * \|\mathcal{L}\|$
l.3 - l.7	$c_2 * n * \|\boldsymbol{\eta}\|$
l.9 - l.15	$c_3 * \|\mathcal{L}\| * (n + c_4)$
l.23 - l.29	$c_5 * \|\boldsymbol{\eta}\|$

$$C_{SP} = n\left[(c_1 + c_3)|\mathcal{L}| + c_2|\boldsymbol{\eta}|\right] + c_3c_4|\mathcal{L}| + c_5|\boldsymbol{\eta}|$$

The assumption of linearity $(c_1, c_2 \cdots)$ is a bit simplifiction and valid only for small ranges of $n, |\mathcal{L}|, |\boldsymbol{\eta}|$. If the ranges are greater, then effects like "paging" of memory, the way files are stored in a concrete file system (e.g. **FAT 32** or **NTFS**) and variables used for storing the "coding" numbers may come in play. E.g. values of $\boldsymbol{\eta}$ are stored in variables of type **integer*1** and therefore should not exceed 255.

Therefore, instead of looking for explicit values for $c_1, \cdots c_5$, direct measurements are documented in Table 1 where length $|\mathcal{L}|$ of $\mathcal{L}$ varies from 1000 to 20000 and in Table 2 where width n of $\mathcal{L}$ varies from 35 to 300. Corresponding files $\mathcal{L}$ (e.g. $\mathcal{L}(70, 1089)$ or $\mathcal{L}(35, 20000)$) were generated from original $\mathcal{L}$ (i.e. $\mathcal{L}(35, 1089)$) by repeating respective rows and columns. In the Tables 1 and 2, column T_{read} contains time necessary to read $\mathcal{L}$. Column T_{total} increases with square power of $|\mathcal{L}|$ as it is the time necessary for $|\mathcal{L}|$ decisions. When T_{total} is divided by $|\mathcal{L}|$, then the times in column $T_{decision}$ are always below 1 sec and therefore completely acceptable. Based both on analysis and direct measurements, complexity of SP is not a problem. Therefore, the limiting factor for better discernment power of SP is an externality i.e. experts should provide bigger data in form of $\mathcal{L}$.

5 Decision-making quality

Though the following example is very simple, it may reveal interesting facts when comparing SP and the decision-making algorithm $A4$(from [5] and mentioned in [6]) that can serve as a simple representative of marginal-based algorithms. Let knowledge base $\mathcal{KB}$ consist of 3 marginals i.e. $\mathcal{KB} = \{m_1, m_2, m_3\}$.

m	T_{read}	T_{total}	$T_{decision}$
1000	31 msec	1 sec	10 msec
2000	60 msec	2 sec	20 msec
5000	156 msec	25 sec	50 msec
10000	343 msec	100 sec	100 msec
20000	687 msec	400 sec	200 msec

Table 1: Computational time dependence on length $m = |\mathcal{F}|$ of file $\mathcal{F}$

n	T_{read}	T_{total}	$T_{decision}$
35	31 msec	0.546 sec	0.5 msec
70	31 msec	1.046 sec	0.9 msec
105	47 msec	1.516 sec	1.3 msec
140	45 msec	1.968 sec	1.8 msec
200	78 msec	2.78 sec	2.59 msec
300	109 msec	4.07 sec	3.78 msec

Table 2: Computational time dependence on width n of file $\mathcal{F}$

The "testing environment" provides an easy way to manipulate with "inputs" to the decision-making algorithm.

First, it makes possible to remove marginals from $\mathcal{KB}$ and second, not all symptom variables, from n possible ones, need to be revealed as "evidences" to decision making-algorithm SP or $A4$.

Then, the expression $\{m_1, m_3\} \cap \{\xi_1 - \xi_{33}\}$ (in column "marginals $\cap$ variables" of Table 3) stands for situation s_2 where $\mathcal{KB}$ consists only of marginals $\{m_1, m_3\}$ and values of all 33 symptom variables $\{\xi_1 - \xi_{33}\}$ are submitted as "evidences" to the SP and $A4$. (Naturally, $\{m_1, m_3\}$ has impact only on $A4$, whereas $\{\xi_1 - \xi_{33}\}$ influences both SP and $A4$). Column "active variables" in Table 3 contains symptom variables whose values have influence on $A4$ as result of both conditions. Column "active space" is product of their ranges. As all symptom variables here are dichotomical ones, the values are like 4, 16, 32.

Let further, the above mentioned marginals describe, beside the implicitly supposed diagnosis variable η, behavior of the following sets of symptom variables:

$$\underline{m_1} = \{\xi_{25}, \xi_{33}\},\ \underline{m_2} = \{\xi_{26}, \xi_{32}\},\ \underline{m_3} = \{\xi_{27}, \xi_{33}\}$$

Even this denotation is a little simplified. E.g. $m_1 = P_{\eta\xi_{25}\xi_{32}}$ is not enough as it should be also mentioned what data was used for populating the marginal m_1. This can be expressed by adding the source. E.g. $m_1 = P_{\eta\xi_{25}\xi_{32}}(\mathcal{L})$ stands for the marginal filled from the data set $\mathcal{L}$. This denotation would do for the column $A4A$, but not for calculating the values for column $A4L$. Then, in fact, there are 1089 different marginals $m_1(\mathcal{L}\backslash t) = P_{\eta\xi_{25}\xi_{32}}(\mathcal{L}\backslash t)$. Those marginals will be populated from 1089 different data files $\mathcal{L}\backslash t)$ that have to be created just for the purpose.

The "A" (in column $A4A$ containing calculation of wrong classifications) stresses that *all* data was used both for learning and testing i.e. $\mathcal{L} = \mathcal{T}$. The "L" (in column $A4L$) has the meaning that the method "Leave one out" was used for the calculation.

It can be observed in Table 4 that L-values are higher than corresponding A-values. In general, SP is slightly better than $A4$, but not always
e.g. $A4L(s8) = 423 < SPL(s8) = 431$. On the basis of other similar experiments, it looks like that advantages of SP may be more prominent but only for A-testing. Especially for $\mathcal{KB}$ with more marginals and when values of all symptom variables are known. As far as "Leave one out" method is concerned and **not** with full-sized evidence, no decisive conclusions can be drawn, so far. However, it seems that SP does quite well and could be used along with other recommended methods.

6 Conclusion

Among positive features of marginal-less SP algorithm, the following ones can be mentioned :

situations	marginals ∩ variables	active variables	active space
s_1	$\{m_1, m_2, m_3\} \cap \{\xi_1 - \xi_{33}\}$	$\xi_{25}, \xi_{26}, \xi_{27}, \xi_{32}, \xi_{33}$	32
s_2	$\{m_1, m_3\} \cap \{\xi_1 - \xi_{33}\}$	$\xi_{25}, \xi_{27}, \xi_{33}$	8
s_3	$\{m_2\} \cap \{\xi_1 - \xi_{33}\}$	ξ_{26}, ξ_{33}	4
s_4	$\{m_1\} \cap \{\xi_1 - \xi_{33}\}$	ξ_{25}, ξ_{33}	4
s_5	$\{m_3\} \cap \{\xi_1 - \xi_{33}\}$	ξ_{27}, ξ_{33}	4
s_6	$\{m_2, m_3\} \cap \{\xi_1 - \xi_{33}\}$	$\xi_{26}, \xi_{27}, \xi_{32}, \xi_{33}$	16
s_7	$\{m_1, m_2, m_3\} \cap \{\xi_1 - \xi_{32}\}$	$\xi_{25}, \xi_{26}, \xi_{27}, \xi_{32}$	16
s_8	$\{m_1, m_2, m_3\} \cap \{\xi_{33}\}$	ξ_{33}	2

Table 3: Different testing situations

situations	SPL	SPA	A4L	A4A
s_1	421	415	427	421
s_2	462	460	463	462
s_3	538	538	538	538
s_4	464	464	464	464
s_5	463	461	463	461
s_6	431	420	423	422
s_7	537	530	539	539
s_8	464	464	464	464

Table 4: Comparing wrong classifications for SP and A4

1. The presented algorithm SP is sufficiently fast i.e. decisions are made within seconds.

2. SP has good discernment power when the tested case t was included in the learning file $\mathcal{L}$ and values of all symptom variables (from $\mathcal{L}$) are given as input evidence.

3. It is easy to add new cases (or remove old ones if considered as obsolete) to the learning file $\mathcal{L}$. In marginal-based approach, it is necessary to recalculate the marginals.

4. Problems associated with selection of marginals are avoided (by definition !) and only symptom variables are necessary. In general, values of all symptom variables (present in the learning file $\mathcal{L}$) should be provided as evidences.

5. Testing via "Leave one out" technique is extremely easy with a small modification in the presented code of SP. It takes approximately the same time as testing on the all data (i.e. when $\mathcal{L} = \mathcal{T}$). Marginal-based algorithm require for "Leave one out" a lot of time for splitting the data file ($|\mathcal{F}|$) times !) and filling the marginals for each split.

SP has several drawbacks as well:

1. SP can be applied only to nominal variables (i.e. not continuous, not cardinal and even not to ordinal).

2. As the only testing criterion is number of wrong classifications, SP is not proper choice for risk analysis.

3. With decreasing number of symptoms (evidences), discernment power of SP drops as well. (It is similar to marginal-based algorithms, as well.)

4. It is not possible to add additional knowledge about structure of P. All is based on input data file $\mathcal{L}$ only.

With respect to above mentioned arguments, SP can be recommended for decision-making on nominal symptom variables and when only a sufficient learning data file is available. It can serve as an alternative to well established marginal-based algorithms for decision-making under uncertainty.

References

[1] P.Cheeseman: A method of computing generalized Bayesian probability values of expert systems with probabilistic background, in: Proc. 6-th Joint Conf. on AI(IJCAI-83), Karlsruhe

[2] W.E. Deming, F.F Stephan: On a least square adjustment of sampled frequency table when expected marginal totals are known, Ann.Math.Stat. 11(1940), pp. 427 - 444

[3] E.T. Jaynes: On the rationale of maximum-entropy methods, Proc. of the IEEE 70 (1980), pp. 939 - 952.

[4] H.G. Kellerer: Verteilungsfunktionen mit gegebenen Marginalverteilungen. Zeitschrift für Wahrscheinlichkeitstheorie, 3(1964), pp. 247 -270.

[5] O. Kříž: 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

[6] O. Kříž: Comparing algorithms based on marginal problem. Kybernetika, Vol 43, (2007), No.5, 633–647

[7] O. Kříž: Selecting marginals for decision making based on marginal problem. WUPES'09, Vol 43, (2009), No.5, 633–647

[8] O. Kříž: Mixing marginals for decision making based on marginal problem. In: WUPES 2012, Proceedings of the 9-th Workshop on Uncertainty Processing, (T. Kroupa,J. Vejnarová ed.), Mariánské Lázně 2012, pp. 114–125.

[9] A.Perez: A probabilistic approach to the integration of partial knowledge for medical decision-making (in Czech). In: Proc. of the 1-st Czechoslovak Congress of Biomedical engineering (BMI83)
(J. Zvárová ed.), Mariánské Lázně 1983, pp. 221–226.

Algorithms for single-fault troubleshooting with dependent actions*

Václav Lín
Department of Decision-Making Theory
Institute of Information Theory and Automation
Czech Academy of Sciences
e-mail: lin@utia.cas.cz

Abstract

We study the problem of single-fault troubleshooting with dependent actions. We propose an integer linear programming formulation for the problem. This can be used to solve the problem directly or to compute lower bounds of optima using linear programming relaxation. We present an optimal dynamic programming algorithm, and three greedy algorithms for computing upper bounds of optima.

1 Introduction

We study *single-fault troubleshooting with dependent actions* [Heckerman et al., 1995, Jensen et al., 2001]. The problem is $\mathcal{NP}$-hard [Vomlelová and Vomlel, 2003], and it is a straightforward generalization of *min-sum set cover* and *pipelined-set cover* [Feige et al., 2004, Munagala et al., 2005]. These are combinatorial problems relevant in several areas other than automated repair.

We propose an integer linear programming formulation for the troubleshooting problem, and give several classes of additional valid inequalities. This can be used to solve the problem directly using a general purpose solver, or to compute lower bounds of optima by linear programming relaxation. We also describe several greedy algorithms for computing upper bounds of optima. We test the resulting lower and upper bounds in computational experiments.

Problem statement The troubleshooting problem studied in this paper may be stated as follows:

- A piece of equipment is faulty and the task is to construct a repair strategy with the least expected cost.

*This work was supported by the Czech Science Foundation through grant 13-20012S.

- There are m mutually exclusive possible causes of the failure called *faults*. The faults are not directly observable. The equipment failure is caused by exactly one of the faults. Each fault F_i has nonzero probability of occurrence $P(F_i)$, and $\sum_{i=1}^{m} P(F_i) = 1$.
- There are n repair steps available, called *actions*, that can possibly remedy the failure. When performed, each action A_j can succeed or fail to fix the system failure, and it has a fixed cost $c(A_j)$ and a conditional probability of success $P(A_j \mid F_i) \geq 0$ for each fault F_i. In terms of probability, the actions are conditionally independent given the faults. It is assumed that an action that has failed once will fail again if performed. Hence, it is assumed that each action is performed at most once.
- The challenge is to find a suitable permutation of the actions $A_1, \ldots, A_n$, and use the permutation as a repair strategy: the actions are performed in the prescribed order until some of the actions succeeds (i.e. the equipment failure is repaired) or all the actions with nonzero probability of success have been used.

Let us denote by $\neg A$ the event that action A has failed and denote by

$$\mathbf{e}_j \quad = \quad \bigwedge_{k=1}^{j} \neg A_{\pi(k)} \ , \tag{1}$$

the information (called *evidence*) that the first j actions in permutation π have failed. Now, for a permutation of actions π we define

$$EC(\pi) \quad = \quad \sum_{i=1}^{n} c(A_{\pi(i)}) \cdot P(\mathbf{e}_{i-1}) \tag{2}$$

$$ECR(\pi) \quad = \quad EC(\pi) - \sum_{\substack{i=1,\ldots,n \\ P(A_{\pi(i)} \mid \mathbf{e}_{i-1})=0}} c(A_{\pi(i)}) \cdot P(\mathbf{e}_{i-1}) \ , \tag{3}$$

where EC is the *expected cost* of π, and ECR is the *expected cost of repair* of π, which is the expected cost of π where the actions with zero probability of success are skipped.

Our task is to find a permutation of actions minimizing the ECR. For some problems, a sequence minimizing EC minimizes also ECR. However, this is not a general rule as shown by Example 1.

Example 1. *We exhibit a troubleshooting problem where there are permutations π_1 and π_2 such that permutation π_1 minimizes ECR and $ECR(\pi_1) < ECR(\pi_2)$, permutation π_2 minimizes EC and $EC(\pi_1) > EC(\pi_2)$. In the problem we have two faults and two actions and the parameters are*

	A_1	A_2
$c(A)$	4	7

	F_1	F_2
$P(F)$	1/2	1/2

	F_1	F_2
$P(A_1 \mid F)$	1	0
$P(A_2 \mid F)$	1	1/2

Let $\pi_1 = \langle A_2, A_1 \rangle$ *and* $\pi_2 = \langle A_1, A_2 \rangle$. *The expected costs are:*

	π_2	π_1
EC	**7.5**	8
ECR	7.5	**7**

Special cases When the outcomes of actions are deterministic, i.e. the probability $P(A \text{ succeeds } \mid F \text{ is present})$ is either zero or one for all combinations of action A and fault F, then the problem reduces to the *pipelined set cover* problem Munagala et al. [2005]. When it is further assumed that all the action cost and fault probabilities are uniform, the problem is equivalent to the *min-sum set cover* problem Feige et al. [2004].

Contribution and structure of the paper The main contribution of the paper is an integer linear programming formulation for single fault troubleshooting with dependent actions. The formulation is useful in two ways:

1. The formulation can be used for solving the troubleshooting problem directly with any general purpose integer programming solver.
2. Even if the integer program at hand is too hard to solve to optimality, we may nonetheless use it to compute lower bounds of optima by linear programming relaxation. These bounds may be used in special purpose branch & bound algorithms for troubleshooting (such as the algorithm given by Vomlelová and Vomlel [2003]).

The integer linear programming formulation is described in Section 5. In Section 4 we describe three simple greedy algorithms for solving the problem. These algorithms are useful because: the search for an optimal permutation of actions by branch & bound algorithms is often greatly facilitated by having a good upper bound of the optimum. The paper concludes with discussion of computational experience in Section 6.

2 Notation

The set of all faults is $\mathcal{F} = \{F_1, \ldots, F_m\}$, the set of all actions is $\mathcal{A} = \{A_1, \ldots, A_n\}$. For an action A, the set of all faults that can be repaired by action A is $\mathcal{F}(A)$; similarly, $\mathcal{A}(F)$ is the set of actions that may repair fault F:

$$\begin{aligned} \mathcal{F}(A) &= \{F \in \mathcal{F} \colon P(A \mid F) > 0\} \ , \\ \mathcal{A}(F) &= \{A \in \mathcal{A} \colon P(A \mid F) > 0\} \ . \end{aligned}$$

Let π be a permutation of the actions. With the notation just introduced and using the assumptions of mutually exclusive faults and conditional independence of actions given faults, the expected cost may be written as

$$EC(\pi) = \sum_{A \in \mathcal{A}} c(A) \cdot \sum_{F \in \mathcal{F}} P(F) \cdot \prod_{\substack{B \in \mathcal{A}(F) \\ \pi(B) < \pi(A)}} P(\neg B | F) \ . \tag{4}$$

When we perform an action A and the action fails, we nevertheless obtain some information. In particular, the marginal probability distribution $P(\mathcal{F})$ changes to $P(\mathcal{F} \mid \neg A)$. As mentioned above, we call this information *evidence*. For consistency, we define $\mathbf{e}_0$, the void *initial evidence* that we have before any of the actions has been executed. We define

$$\mathcal{A}(\mathbf{e}) \;=\; \{A \in \mathcal{A}\colon P(A \mid \mathbf{e}) > 0\}\ .$$

It is assumed that once failed action will fail again if performed. In terms of probability, $P(A \mid \neg A)$ is zero, and hence the set $\mathcal{A}(\mathbf{e})$ does not contain any of the actions that are included in $\mathbf{e}$.

3 Dynamic programming

A dynamic programming approach to troubleshooting was first proposed by Vomlelová and Vomlel [2003]. The problem studied in this paper can be solved by dynamic programming using recurrence

$$ECR^\star(\mathbf{e}) \quad = \quad \min_{A \in \mathcal{A}(\mathbf{e})} \left[c'(A \mid \mathbf{e}) + P(\neg A \mid \mathbf{e}) \cdot ECR^\star(\mathbf{e} \wedge \neg A)\right]\ , \tag{5}$$

where

$$c'(A \mid \mathbf{e}) \quad = \quad \begin{cases} 0 & \text{if } P(\neg A \mid \mathbf{e}) = 1 \\ c(A) & \text{otherwise} \end{cases}\ .$$

Now, $ECR^\star(\mathbf{e}_0)$ is the optimal expected cost of the troubleshooting problem.

4 Greedy algorithms

Greedy polynomial-time algorithms may be used to construct permutations of actions that are not guaranteed to be optimal but experience shows that very often they are optimal or "nearly optimal". We shall describe three such algorithms in this section.

Algorithm Updating P/C Perhaps the most natural greedy algorithm is one called Updating P/C [Jensen et al., 2001] At i^{th} step, $i = 1, \ldots, n$, the algorithm selects an action $A \in \mathcal{A}(\mathbf{e}_{i-1})$ maximizing the ratio

$$\frac{P(A \mid \mathbf{e}_{i-1})}{c(A)}\ .$$

For problems where EC and ECR are minimized by the same permutation, Kaplan et al. [2005] proved that Updating P/C has a guaranteed approximation factor: it never returns a sequence with expected cost greater than four times the optimum. By the complexity-theoretic results of Feige et al. [2004], that is most likely the best guaranteed approximation factor possible.

Algorithm DP-greedy Another greedy algorithm is motivated by the dynamic programming recurrence (5). We shall call the algorithm DP-GREEDY. At i^{th} step, the algorithm selects an action $A \in \mathcal{A}(\mathbf{e}_{i-1})$ minimizing

$$c(A) + P(\neg A \mid \mathbf{e}_{i-1}) \cdot \widetilde{EC}(\mathbf{e}_{i-1} \wedge \neg A) \ ,$$

where $\widetilde{EC}(\mathbf{e})$ denotes an estimate of the expected cost of optimal sequence of the remaining actions from $\mathcal{A}(\mathbf{e})$. The estimate is computed by the UPDATING P/C algorithm. Although seemingly different, algorithm DP-GREEDY is equivalent to a greedy algorithm proposed by Langseth and Jensen [2001].[1]

Algorithm I-greedy The last greedy algorithm uses an information-theoretic criterion for selection of the hopefully best action given evidence $\mathbf{e}$. We call it I-GREEDY. It is inspired by the ID3 algorithm [Quinlan, 1986]. Let $H(\mathcal{F} \mid \mathbf{e})$ be the Shannon entropy of marginal distribution $P(\mathcal{F} \mid \mathbf{e})$, that is

$$H(\mathcal{F} \mid \mathbf{e}) \quad = \quad \sum_{F \in \mathcal{F}} P(F \mid \mathbf{e}) \cdot \log \frac{1}{P(F \mid \mathbf{e})} \ ,$$

and let $I(A \mid \mathbf{e})$ be the *information gain* of performing action A, i.e. the expected decrease of $H(\mathcal{F} \mid \mathbf{e})$ induced by performing A:

$$I(A \mid \mathbf{e}) \quad = \quad H(\mathcal{F} \mid \mathbf{e}) - \Big[P(A \mid \mathbf{e}) \cdot H(\mathcal{F} \mid \mathbf{e} \wedge A) + P(\neg A \mid \mathbf{e}) \cdot H(\mathcal{F} \mid \mathbf{e} \wedge \neg A) \Big].$$

Given evidence $\mathbf{e}$, it seems desirable to select an action maximizing ${}^{I(A \mid \mathbf{e})}/_{c(A)}$. However, we do not want the selection criterion to be biased towards actions with high entropy $H(\mathcal{F} \mid \mathbf{e} \wedge A)$ since we are not interested in the entropy of $P(\mathcal{F})$ in the case that A suceeds. To this end, we assume $H(\mathcal{F} \mid \mathbf{e} \wedge A)$ to be zero for all A and $\mathbf{e}$, and we select at each step an action $A \in \mathcal{A}(\mathbf{e}_{i-1})$ maximizing

$$\frac{H(\mathcal{F} \mid \mathbf{e}) - P(\neg A \mid \mathbf{e}_{i-1}) \cdot H(\mathcal{F} \mid \mathbf{e}_{i-1} \wedge \neg A)}{c(A)} \ . \tag{6}$$

5 Integer linear programming formulation

We shall formulate an integer linear program encoding the troubleshooting problem. For background information about integer programming we refer to [Wolsey, 1998]. For linear programming, the classic reference is [Dantzig, 1998].

To encode a permutation of actions from the set $\mathcal{A}$, we use binary variables $d_{A,B}$ for every pair of distinct actions $A, B \in \mathcal{A}$. Given a permutation π of the actions, we have $d_{A,B} = 1$ if action A precedes action B in the permutation π, otherwise $d_{A,B} = 0$. Variables $d_{A,B}$ should encode a linear ordering relation on $\mathcal{A}$. That means that the relation is *asymmetric* and *transitive* To enforce the requirement of asymmetry, we introduce equation (7) for each ordered pair of distinct actions A, B. To enforce the

[1] The proof is to be found in [Lín, 2015].

requirement of transitivity, we add inequality (8) for every ordered triple of pairwise distict actions A, B, C.

$$d_{A,B} = 1 - d_{B,A} \ . \tag{7}$$

$$d_{A,B} + d_{B,C} \leq d_{A,C} + 1 \ . \tag{8}$$

We now proceed to formulate the expected cost of action sequence as a linear function. For simplicity, we begin by EC and turn to ECR later. Assuming that a fixed permutation π is encoded by variables $d_{A,B}$ introduced above, we can write (4) as:

$$\sum_{A \in \mathcal{A}} c(A) \cdot \sum_{F \in \mathcal{F}} P(F) \cdot \prod_{\substack{B \in \mathcal{A}(F) \setminus \{A\} \\ d_{B,A}=1}} P(\neg B \mid F) \ . \tag{9}$$

(Whenever the product in (9) is taken over an empty set of factors, we assume that the product equals one. That is $\prod_{B \in \emptyset} P(\neg B \mid F) = 1$.) Minimizing (4) is equivalent to minimizing (9) subject to the constraints (7) and (8). To express (9) as a linear function, we introduce a binary variable $x_{F,A,\mathcal{B}}$ for each fixed combination of fault F, action A and a set of actions $\mathcal{B} \subseteq \mathcal{A}(F) \setminus \{A\}$. The value of $x_{F,A,\mathcal{B}}$ is defined as

$$x_{F,A,\mathcal{B}} = \Big(\prod_{B \in \mathcal{B}} d_{B,A} \Big) \cdot \Big(\prod_{\substack{B \in \mathcal{A}(F) \setminus \mathcal{B} \\ B \neq A}} d_{A,B} \Big) \ . \tag{10}$$

In words, variable $x_{F,A,\mathcal{B}}$ equals one if and only if all the actions $B \in \mathcal{B}$ precede action A, and all the remaining actions from $\mathcal{A}(F)$ are preceded by A. Associated to each variable $x_{F,A,\mathcal{B}}$ is a coefficient

$$Q_{F,A,\mathcal{B}} = c(A) \cdot P(F) \cdot \prod_{B \in \mathcal{B}} P(\neg B \mid F) \ .$$

For $\mathcal{B} = \emptyset$, we have $Q_{F,A,\mathcal{B}} = c(A) \cdot P(F)$. For any fixed fault F and action A, exactly one of the variables $x_{F,A,\mathcal{B}}$ equals one. With this observation, we may replace the nonlinear objective (9) by a linear function

$$EC = \sum_{A \in \mathcal{A}} \sum_{F \in \mathcal{F}} \sum_{\substack{\mathcal{B} \subseteq \mathcal{A}(F) \\ \mathcal{B} \not\ni A}} Q_{F,A,\mathcal{B}} \cdot x_{F,A,\mathcal{B}} \ . \tag{11}$$

The number of summands in (11) is exponential in the size of sets $\mathcal{A}(F)$. However, we can assume that in practical applications, the size $|\mathcal{A}(F)|$ is bounded from above by a reasonably small constant.

To express definition (10) in terms of linear inequalities, we observe that the definition implies for each fixed combination $F, A, \mathcal{B}$:

$$x_{F,A,\mathcal{B}} \geq 1 + \sum_{B \in \mathcal{B}} (d_{B,A} - 1) + \sum_{\substack{B \in \mathcal{A}(F) \setminus \mathcal{B} \\ B \neq A}} (d_{A,B} - 1). \tag{12}$$

Bounding the variables $x_{F,A,\mathcal{B}}$ from above is not necessary since all the coefficients in (11) are nonegative. In case that $\mathcal{A}(F) \setminus \{A\}$ is an empty set, we have $x_{F,A,\emptyset} = 1$. To summarize, for minimization of EC we have a minimization linear program with objective function (11) and constraints (12), (7), (8).

To minimize ECR rather than EC, we need to add to the formulation additional variables and constraints. We say that an action A_i is *dominated* in permutation π if its success probability $P(A_{\pi(i)} \mid \mathbf{e}_{i-1})$ is zero. In the linear model, we define additional variable $w_{F,A,\mathcal{B}}$ with coefficient '$-Q_{F,A,\mathcal{B}}$' whenever the following conditions hold:

1. $(\forall G \in \mathcal{F}(A))(\exists B \in \mathcal{A}(G) \setminus \{A\}) \quad P(B \mid F) = 1$, i.e. action A can be dominated, and
2. $(\forall B \in \mathcal{B}) \quad P(B \mid F) < 1$, i.e. the coefficient $Q_{F,A,\mathcal{B}}$ is nonzero,

We do not create variables $w_{F,A,\mathcal{B}}$ where $\mathcal{B} = \emptyset$ and $A \in \mathcal{A}(F)$ since $\mathcal{B} = \emptyset$ means that action A is not preceded by any action $B \in \mathcal{A}(F)$ and hence A cannot be dominated. Each variable $w_{F,A,\mathcal{B}}$ has to equal one if and only if the corresponding variable $x_{F,A,\mathcal{B}}$ equals one, and the action A is dominated. To express this requirement by linear inequalities, we observe that only upper bound for the w-variables is needed (since the coefficients of w are negative), and it is sufficient to introduce linear constraints

$$w_{F,A,\mathcal{B}} \leq x_{F,A,\mathcal{B}} \tag{13}$$

$$(\forall G \in \mathcal{F}(A)) \quad w_{F,A,\mathcal{B}} \leq \sum_{B \in \mathcal{D}(A,G)} d_{B,A} \tag{14}$$

where

$$\mathcal{D}(A,G) = \{B \in \mathcal{A}(G) \setminus \{A\} \colon P(B \mid G) = 1\} .$$

The linear objective function is then

$$ECR = \sum_{x_{F,A,\mathcal{B}}} Q_{F,A,\mathcal{B}} \cdot x_{F,A,\mathcal{B}} - \sum_{w_{F,A,\mathcal{B}}} Q_{F,A,\mathcal{B}} \cdot w_{F,A,\mathcal{B}} . \tag{15}$$

5.1 Classes of additional valid inequalities

Linear programming relaxation is obtained from the integer program by replacing the integrality requirement $d_{A,B} \in \{0,1\}$ by $0 \leq d_{A,B} \leq 1$ for all the d-variables and likewise for all the x- and w-variables. In general, objective value of linear programming relaxation is a lower bound of the objective value of the minimization integer program. To make the bound as tight as possible, we may add to the linear model additional *valid inequalities*. That is, inequalities that are satisfied by all feasible integer solutions.

The first class of valid inequalities is based on the observation that for any fixed combination of a fault F and an action A, exactly one of the variables $x_{F,A,\mathcal{B}}$ equals one, i.e.

$$\sum_{\mathcal{B} \subseteq \mathcal{A}(F) \setminus \{A\}} x_{F,A,\mathcal{B}} = 1 . \tag{16}$$

By (13), each equality (16) induces a corresponding inequality over the w-variables:

$$\sum_{\mathcal{B}} w_{F,A,\mathcal{B}} \quad \leq \quad 1 \ . \tag{17}$$

where the left hand sum is taken over all the w-variables (if any) with appropriate indices F and A.

Another class of valid inequalities is based on observing that given fault F and a fixed permutation of actions, there is exactly one action $A \in \mathcal{A}(F)$ that is not preceded by any other action $B \in \mathcal{A}(F)$. That is, for every fault F we have:

$$\sum_{A \in \mathcal{A}(F)} x_{F,A,\emptyset} \quad = \quad 1 \ . \tag{18}$$

We observe that if action B precedes action A, and action A precedes all the actions from $\mathcal{A}(F)$, then also action B precedes all the actions from $\mathcal{A}(F)$. Hence, for any fixed combination of fault F and distinct actions A and B we have:

$$x_{F,A,\emptyset} + d_{B,A} \quad \leq \quad x_{F,B,\emptyset} + 1 \ . \tag{19}$$

Another idea for valid inequalities is based on the fact that if an action is dominated in optimal sequence, then so should be its successors. Therefore, for all distinct actions A and B neither of which belongs to $\mathcal{A}(F)$, we have

$$w_{F,A,\emptyset} + d_{A,B} \quad \leq \quad w_{F,B,\emptyset} + 1 \ . \tag{20}$$

For any fixed triple $F, A, \mathcal{B}$, a combination of (10) and (13) yields inequalities

$$w_{F,A,\mathcal{B}} \quad \leq \quad d_{B,A} \text{ for every } B \in \mathcal{B} \tag{21}$$

$$w_{F,A,\mathcal{B}} \quad \leq \quad d_{A,B} \text{ for every } B \in \mathcal{A}(F) \setminus \mathcal{B} \setminus \{A\} \ . \tag{22}$$

Another class of valid inequalities that we devise is inspired by a heuristic function due to Vomlelová and Vomlel [2003]. For any given fault F we can find a permutation π_F of all the actions minimizing

$$z(\pi) \quad = \quad \sum_{A \in \mathcal{A}} c(A) \cdot \prod_{\substack{B \in \mathcal{A} \\ \pi(B) < \pi(A)}} P(\neg B \mid F) \ .$$

The minimizing permutation π_F is found by ordering the actions in $\mathcal{A}$ so that the ratios $P(A \mid F)/c_A$ are nonincreasing. With this observation, we can construct constraints (23) for each fault F:

$$\sum_{A \in \mathcal{A}} \sum_{\mathcal{B}} Q_{F,A,\mathcal{B}} \cdot x_{F,A,\mathcal{B}} - \sum_{A \in \mathcal{A}(F)} \sum_{\mathcal{B}} Q_{F,A,\mathcal{B}} \cdot w_{F,A,\mathcal{B}} \quad \geq \quad P(F) \cdot z(\pi_F) \ . \tag{23}$$

The heuristic function of Vomlelová and Vomlel [2003] can be expressed by a single inequality:

$$\sum_{x_{F,A,\mathcal{B}}} Q_{F,A,\mathcal{B}} \cdot x_{F,A,\mathcal{B}} - \sum_{w_{F,A,\mathcal{B}}} Q_{F,A,\mathcal{B}} \cdot w_{F,A,\mathcal{B}} \quad \geq \quad \sum_{F \in \mathcal{F}} P(F) \cdot z(\pi_F) \ . \tag{24}$$

The sums in (24) are taken over all the x- and w-variables that exist in the linear programming formulation.

Fixing partial order of actions in advance In some cases, we can fix a partial order of some of the actions before starting to search for an optimal sequence, thereby reducing the number of sequences that need to be considered. In particular, we may use the following proposition.[2]

Proposition 1. *Let* $\mathbf{s}$ *be an optimal sequence of actions. Let there be two distinct actions* A *and* B *in* $\mathbf{s}$ *such that:*

- *the sets* $\mathcal{F}(A)$ *and* $\mathcal{F}(B)$ *are disjoint,*
- *there is no action* C *with set* $\mathcal{F}(C)$ *insersecting* $\mathcal{F}(A) \cup \mathcal{F}(B)$.

Further, assume that action A *precedes action* B *in sequence* $\mathbf{s}$ *(the two actions are not necessarily adjacent). Then* ${}^{P(A)}/_{c(A)} \geq {}^{P(B)}/_{c(B)}$.

We may use Proposition 1 to construct a partial ordering of actions satisfying the conditions stated in the proposition. Once such an ordering is constructed, we may fix the corresponding precedence variables $d_{A,B}$ to appropriate values.

Cutting planes procedure The basic integer programming formulation contains inequalities (12),(7),(8), (13) and (14). The formulation may be strenghtened by adding additional valid equalities and inequalities mentioned above. However for computational reasons, we do not add them all at once, but rather in an iterative fashion. The additional constraints are conventionally called *cutting planes*. The procedure of adding cutting planes can be outlined as follows:

1. Construct the initial integer programming formulation and compute its relaxation by replacing the integrality requirements $v \in \{0, 1\}$ by $0 \leq v \leq 1$ for every variable v of the formulation.
   Let X denote the obtained linear programming formulation, and let $\mathbf{x}$ denote its solution vector found by linear programming.
2. For each class of valid inequalities or equalities[3] listed in Section 5.1:
   (a) Investigate whether some inequalities of the class are violated by the current solution vector $\mathbf{x}$.
   (b) Add the violated inequalities to X and solve by linear programming.
   (c) If the addition of violated inequalities lead to increase in the objective value, keep the inequalities in X. Otherwise, remove them.
   (d) Remove from the formulation cutting planes that are satisfied but have nonzero slack value.
3. Repeat the previous step until the objective value does not increase, or the number of iterations exceeds some predetermined parameter, or $\mathbf{x}$ is an integral vector.

A more detailed description of the procedure is to appear in [Lín, 2015].

[2] Proposition 1 is a straightforward generalization of a theorem proved by Jensen et al. [2001]. The proof is to be found in [Lín, 2015].

[3] In the following, we say just "inequalities" instead of "inequalities or equalities".

6 Computational experience

In this last section we collect results of a small computational study. The algorithms described in the paper were run on nine problems. One of the problems was generated, the other were extracted from real world troubleshooting models. Full details of the models cannot be given for confidentiality reasons, so we provide only some basic characteristics in Table 1. More details can be provided upon request.

		$c(A)$			$P(F)$		$P(A \mid F) \neq 0$		
prob.	$\lvert\mathcal{A}\rvert$	μ	σ	$\lvert\mathcal{F}\rvert$	μ	σ	%	μ	σ
P1	25	6,840	3,923	26	0,038	0,037	8,800	0,886	0,125
P2	13	24,231	30,868	12	0,083	0,055	9,600	0,941	0,066
P3	7	10,429	11,588	6	0,167	0,158	23,800	0,898	0,175
P4	13	28,154	31,945	13	0,077	0,093	15,380	0,950	0,050
P5	10	1,000	0,000	10	0,100	0,000	27,000	0,929	0,051
P6	14	83,000	264,406	13	0,077	0,056	24,720	0,931	0,092
P7	20	10,900	7,840	26	0,039	0,033	6,920	0,930	0,200
P8	13	34,690	40,400	12	0,083	0,078	20,510	0,963	0,092
P9	11	26,450	35,708	11	0,091	0,118	15,700	0,935	0,068

Table 1: For each problem, we give the number of actions $|\mathcal{A}|$ and number of faults $|\mathcal{F}|$. We also give mean μ and standard deviation σ for action costs $c(A)$ and fault probabilities $P(F)$. For probability distribution $P(\mathcal{A} \mid \mathcal{F})$ we give the percentage (%), mean and standard deviation of nonzero entries.

We investigate tightness of the upper bounds computed by greedy algorithms DP-GREEDY,UPDATING P/C and I-GREEDY. The tightness is measured by ratio of the upper bound to the optimal ECR. The optima are computed by dynamic programming and/or by solving the integer programming formulation. The results are in Table 2. In the same table are ratios of lower bounds to optimal ECR. The lower bounds are computed by the heuristic function of Vomlelová and Vomlel [2003] (column "heur."), by linear programming relaxation without cutting planes (column "LP") and by linear programming relaxation with cutting planes involved (column "LP w. cuts"). In Table 3 we give similar results for the lower bounds when EC is optimized rather than ECR.

We see that the greedy algorithms provide solutions that are always either optimal or very close to optimal. The algorithm DP-GREEDY performs very well and finds an optimal solution in most cases[4]. As far as the lower bounds are concerned, adding cutting planes to the basic linear programming formulation leads to a tighter bound in most cases. In general, lower bounds computed by the cutting planes procedure are the strongest. In some cases however, they are no better than the bounds provided by the simple heuristic proposed by Vomlelová and Vomlel [2003]. We also note that the linear programming relaxation provides tighter bounds when optimizing EC rather than ECR.

[4]Without providing a proof that the solution is in fact optimal.

prob.	DP-GR.	UPD. P/C	I-GR.	heur.	LP	LP w. cuts
P1	1,000	1,000	1,003	0,590	0,408	0,624
P2	1,005	1,006	1,018	0,563	0,973	0,990
P3	1,000	1,000	1,000	0,867	0,796	0,867
P4	1,000	1,000	1,024	0,619	0,742	0,861
P5	1,000	1,047	1,047	0,436	0,379	0,574
P6	1,000	1,000	1,022	0,856	0,882	0,886
P7	1,001	1,010	1,014	0,648	0,909	0,918
P8	1,000	1,000	1,000	0,630	0,334	0,630
P9	1,000	1,000	1,019	0,699	0,861	0,923

Table 2: Tightness of upper and lower bounds of ECR. In the columns on the left are ratios upper bounds to optimal ECR. In the columns on the right are ratios of lower bounds to optimal ECR.

prob	heur.	LP	LP w. cuts
P1	0,577	0,564	0,829
P2	0,563	0,973	0,990
P3	0,818	0,787	0,950
P4	0,619	0,742	0,861
P5	0,436	0,379	0,574
P6	0,856	0,918	0,996
P7	0,609	0,983	0,983
P8	0,619	0,539	0,828
P9	0,699	0,861	0,923

Table 3: Tightness of lower bounds of EC: ratios of lower bounds to optimal EC. The meaning of column names is as in Table 2.

Acknowledgement

I thank Jiří Vomlel and Thorsten Ottosen for their help.

References

George B. Dantzig. *Linear programming and extensions.* Princeton University Press, 1998.

Uriel Feige, László Lovász, and Prasad Tetali. Approximating Min Sum Set Cover. *Algorithmica*, 40(4):219–234, 2004.

David Heckerman, John S. Breese, and Koos Rommelse. Decision-theoretic troubleshooting. *Communications of the ACM*, 38(3):49–57, 1995.

Finn Verner Jensen, Uffe Kjærulff, Brian Kristiansen, Helge Langseth, Claus Skaanning, Jiří Vomlel, and Marta Vomlelová. The SACSO methodology for troubleshooting complex systems. *AI EDAM*, 15(4):321–333, 2001.

Haim Kaplan, Eyal Kushilevitz, and Yishay Mansour. Learning with attribute costs. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 356–365. ACM, 2005.

Helge Langseth and Finn Verner Jensen. Heuristics for two extensions of basic troubleshooting. In *IN: Seventh Scandinavian conference on Artificial Intelligence, SCAI'01, Frontiers in Artificial Intelligence and applications, IOS*, 2001.

Václav Lín. Forthcoming thesis, 2015.

Kamesh Munagala, Shivnath Babu, Rajeev Motwani, and Jennifer Widom. The pipelined set cover problem. In Thomas Eiter and Leonid Libkin, editors, *ICDT*, volume 3363 of *Lecture Notes in Computer Science*, pages 83–98. Springer, 2005. ISBN 3-540-24288-0.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Marta Vomlelová and Jiří Vomlel. Troubleshooting: NP-hardness and solution methods. *Soft Computing*, 7(5):357–368, 2003.

Laurence A. Wolsey. *Integer programming.* Wiley, New York, 1998.

Hierarchical Models as Marginals of Hierarchical Models

Guido Montúfar
Max Planck Institute for
Mathematics in the Sciences
montufar@mis.mpg.de

Johannes Rauh
Department of Mathematics and Statistics
York University
jarauh@yorku.ca

Abstract

We investigate the representation of hierarchical models in terms of marginals of other hierarchical models with smaller interactions. We focus on binary variables and marginals of pairwise interaction models whose hidden variables are conditionally independent given the visible variables. In this case the problem is equivalent to the representation of linear subspaces of polynomials by feedforward neural networks with soft-plus computational units. We show that any binary hierarchical model with M pure higher order interactions can be expressed as the marginal of a pairwise interaction model with $\sim \frac{1}{2}M$ hidden binary variables.

1 Introduction

Consider a finite set V of random variables. A hierarchical log-linear model is a set of joint probability distributions that can be written as products of interaction potentials, as $p(x) = \prod_\Lambda \psi_\Lambda(x)$, where $\psi_\Lambda(x) = \psi_\Lambda(x_\Lambda)$ only depends on the subset Λ of variables and where the product runs over a fixed family of sets Λ. By introducing hidden variables, it is possible to express the same probability distributions in terms of potentials which involve only small sets of variables, as $p(x) = \sum_y \prod_\lambda \psi_\lambda(x, y)$, with small sets λ. Using small interactions is a central idea in the context of connectionistic models, where the sets λ are often restricted to have cardinality two. Due to the simplicity of their local characteristics, these models are particularly well suited for Gibbs sampling [1]. The representation, or explanation, of complex interactions among observed variables in terms of hidden variables is also related to the study of common ancestors [7].

We are interested in sufficient and necessary conditions on the number of hidden variables, their values, and the interaction structures under which a given hierarchical model can be represented as the visible marginal of another hierarchical model with hidden variables. In this work, we focus on binary visible and hidden variables. For the hierarchical models with hidden variables, we restrict our attention to models

involving only pairwise interactions and whose hidden variables are conditionally independent given the visible variables (that is, there are no interactions among the hidden variables). The free energy function of such a model is a sum of soft-plus computational units $x \mapsto \log(1+\exp(\sum_{i\in V} w_i x_i + c))$. On the other hand, the energy function of a fully observable hierarchical model with binary variables is a polynomial, with the monomials corresponding to the pure interactions. Observing that any function that depends on binary variables can be expressed as a polynomial, the task is then to characterize the polynomials computable by a soft-plus unit.

Using this approach, Younes [8] showed that a hierarchical model with N binary variables and a total of M pure higher order interactions (among three or more variables) can be represented as the visible marginal of a pairwise interaction model with M hidden binary variables. In Younes' construction, each pure interaction between a set of visible variables of the original model is modeled by one hidden binary variable that interacts pairwise with each of the involved visible variables. In fact this replacement can be accomplished without increasing the number of model parameters, by imposing linear constraints on the coupling strengths of the hidden variable [8]. In this work, we investigate ways of squeezing more degrees of freedom out of each hidden variable. An indication that this should be possible is the fact that the full interaction model, for which $M = 2^N - \binom{N}{2} - N - 1$, can be modeled with $2^{N-1} - 1$ hidden variables [4]. Indeed, by controlling two polynomial coefficients at the time, we show that in general $\sim \frac{1}{2}M$ hidden variables are sufficient.

A special case of hierarchical models with hidden variables are mixtures of hierarchical models. The smallest mixtures of hierarchical models that contain other hierarchical models have been studied in [3]. For the necessary conditions, the idea there is to compare the possible support sets of the limit distributions of both models. For the sufficient conditions, the idea is to find a small S-set covering of the set of elementary events. An S-set of a probability model is a set of elementary events such that every distribution supported in that set is a limit distribution from the model. Mixture models are closely related to tree models. The geometry of binary tree models was studied in [9] in terms of moments and cumulants via Möbius inversions.

This paper is organized as follows. Section 2 introduces hierarchical models, formalizes and motivates the problem in the light of previous results. Section 3 pursues a characterization of the polynomials that can be represented by soft-plus units. Section 4 applies the obtained characterization to study the representation of hierarchical models in terms of pairwise interaction models, especially restricted Boltzmann machines. Section 5 discusses open problems.

2 Preliminaries

This section introduces hierarchical models with and without hidden variables, formalizes the problem and presents motivating prior results.

2.1 Hierarchical Models

Consider a finite set V of variables with joint states $x = (x_i)_{i\in V} \in \mathbb{X} = \times_{i\in V}\mathbb{X}_i$. For a given set $S \subseteq 2^V$ of subsets of V let

$$\mathcal{V}_{\mathbb{X},S} := \left\{ f(x) = \sum_{\Lambda\in S} f_\Lambda(x) \colon f_\Lambda(x) = f_\Lambda(x_\Lambda) \right\}.$$

This is the linear subspace of $\mathbb{R}^{\mathbb{X}}$ spanned by functions f_Λ that only depend on sets of variables $\Lambda \in S$. The hierarchical model of probability distributions on $\mathbb{X}$ with interactions S is the set

$$\mathcal{E}_{\mathbb{X},S} := \left\{ p(x) = \frac{1}{Z(f)} \exp(f(x)) \colon f \in \mathcal{V}_{\mathbb{X},S} \right\}, \tag{1}$$

where $Z(f) = \sum_{x'\in\mathbb{X}} \exp(f(x'))$ is a normalizing factor. The energy function of a probability distribution from $\mathcal{E}_{\mathbb{X},S}$ is given by

$$E(x) = \sum_{\Lambda\in S} f_\Lambda(x). \tag{2}$$

For convenience, in all what follows we assume that S is a simplicial complex, meaning that $A \in S$ implies $B \in S$ for all $B \subseteq A$. Furthermore, we assume that the union of elements of S equals V. In the case of binary variables the energy can be written as a polynomial, as

$$E(x) = \sum_{\Lambda\in S} J_\Lambda \prod_{i\in\Lambda} x_i.$$

Here, $J_\Lambda \in \mathbb{R}$, $\Lambda \in S$, are the interaction weights that parametrize the model.

2.2 Hierarchical Models with Hidden Variables

Consider an additional set H of variables, with joint states $y = (y_j)_{j\in H} \in \mathbb{Y} = \times_{j\in H}\mathbb{Y}_j$. For a simplicial complex $T \subseteq 2^{V\cup H}$, let $\mathcal{V}_{\mathbb{X}\times\mathbb{Y},T} \subseteq \mathbb{R}^{\mathbb{X}\times\mathbb{Y}}$ be the linear subspace of functions of the form $g(x,y) = \sum_{\lambda\in T} g_\lambda(x,y)$, $g_\lambda(x,y) = g_\lambda((x,y)_\lambda)$. The marginal on $\mathbb{X}$ of the hierarchical model $\mathcal{E}_{\mathbb{X}\times\mathbb{Y},T}$ is the set

$$\mathcal{M}_{\mathbb{X}\times\mathbb{Y},T} := \left\{ p(x) = \frac{1}{Z(g)} \sum_{y\in\mathbb{Y}} \exp(g(x,y)) \colon g \in \mathcal{V}_{\mathbb{X}\times\mathbb{Y},T} \right\}, \tag{3}$$

where $Z(g) = \sum_{x'\in\mathbb{X},y'\in\mathbb{Y}} \exp(g(x',y'))$ is a normalizing factor. The free energy of a probability distribution from $\mathcal{M}_{\mathbb{X}\times\mathbb{Y},T}$ is given by

$$F(x) = \log \sum_{y\in\mathbb{Y}} \exp\Big(\sum_{\lambda\in T} g_\lambda(x,y) \Big). \tag{4}$$

If there are no interactions between hidden variables, i.e. if $|\lambda \cap H| \leq 1$, then this rewrites to

$$F(x) = \sum_{\lambda : \lambda \cap H = \emptyset} g_\lambda(x) + \sum_{j \in H} \log \sum_{y_j \in \mathbb{Y}_j} \exp\Big(\sum_{\lambda \in T : j \in \lambda} g_\lambda(x, y_i) \Big). \tag{5}$$

Particularly interesting are the models with full bipartite interactions between the set of visible variables and the set of hidden variables, $T = \{\lambda \subseteq V \cup H \colon |\lambda \cap V| \leq 1, |\lambda \cap H| \leq 1\}$, called restricted Boltzmann machines.

In the case of binary visible variables (and arbitrary interactions), the free energy can be written as a polynomial, as

$$F(x) = \sum_{B \subseteq V} K_B \prod_{i \in B} x_i,$$

where the coefficients can be computed from Möbius inversion formula as

$$K_B = \sum_{C \subseteq B} (-1)^{|B \setminus C|} \log \sum_{y \in \mathbb{Y}} \exp\Big(\sum_{\lambda \in T} g_\lambda((1_C, 0_{V \setminus C}), y) \Big), \quad B \subseteq V. \tag{6}$$

Here $(1_C, 0_{V \setminus C}) \in \{0,1\}^V$ is the vector with value 1 in the entries $i \in C$ and value 0 in the entries $i \notin C$.

In most cases the marginal of a hierarchical model is itself not a hierarchical model. However, one may ask which hierarchical models are contained in the marginal of a hierarchical model.

2.3 Problem and Previous Results

To represent a hierarchical model in terms of the marginal of another hierarchical model, we need to represent (1) in terms of (3). Equivalently, we need to represent (2) in terms of (4). Given a set of visible variables V and a simplicial complex $S \subseteq 2^V$, what conditions on the set of hidden variables H and the simplicial complex $T \subseteq 2^{V \cup H}$ are sufficient and necessary in order for any function E of the form (2) to be representable in terms of some function F of the form (4)? We would like to arrive at a result that generalizes the following.

- A restricted Boltzmann machine with $|H|$ hidden binary variables can approximate any probability distribution from any binary hierarchical model $\mathcal{E}_S$ with $|S \setminus \binom{V}{1}| - 1 \leq |H|$ arbitrarily well. See [8].

- The restricted Boltzmann machine with $|H| = 2^{|V|-1} - 1$ hidden binary variables can approximate any probability distribution on $\{0,1\}^V$ arbitrarily well. See [4].

- Every probability distribution on $\{0,1\}^V$ can be approximated arbitrarily well by some mixture of k fully factorizing probability distributions if and only if $k \geq 2^{|V|-1}$. See [3].

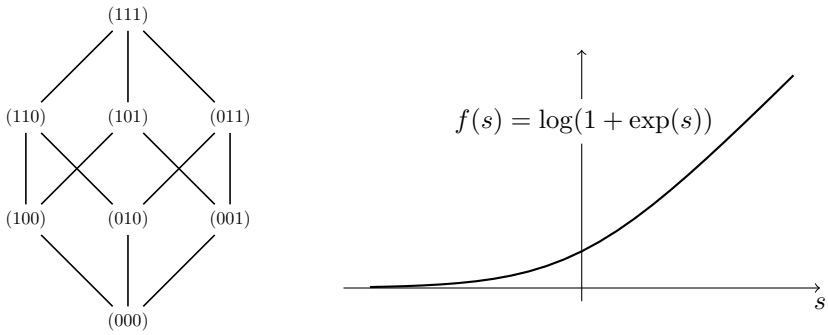


Figure 1: Illustration of a soft-plus computational unit. The possible inputs, corresponding to the vertices of a cube, are mapped to the real line by an affine map, and then the soft-plus non-linearity $s \mapsto \log(1 + \exp(s))$ is applied.

Our Theorem 5 below improves the first item and almost recovers the second item for the special case of approximating the set of all probability distributions. The third item is an example of a tight bound, providing sufficient and necessary conditions. The set of mixtures of k fully factorizing probability distributions corresponds to the hierarchical model with one k-valued hidden variable that interacts pairwise with each visible variable.

3 Soft-plus Polynomials

Consider the functions of the form $\phi\colon \{0,1\}^V \to \mathbb{R};\ x \mapsto \log(1 + \exp(w^\top x + c))$, parametrized by $w \in \mathbb{R}^V$ and $c \in \mathbb{R}$. This corresponds to the free energy added by one hidden binary variable interacting pairwise with each visible binary variable; see Equation (5). We regard ϕ as a *soft-plus computational unit*, which integrates an input vector x into a scalar via $x \mapsto w^\top x + c$, and applies the soft-plus non-linearity $s \mapsto \log(1+\exp(s))$. See Figure 1. What polynomials can be represented in this way? Following Equation (6), the polynomial coefficients of ϕ are given by

$$K_B(w,c) = \sum_{C \subseteq B} (-1)^{|B \setminus C|} \log\left(1 + \exp\Big(\sum_{i \in C} w_i + c\Big)\right), \quad B \in 2^V.$$

This is an alternating sum of the values of the soft-plus unit on the input vectors with $\mathrm{supp}(x) \subseteq B$.

The monomials of partial degree one are partially ordered by inclusion, as illustrated in Figure 2. We focus on the description of the possible values of the highest degree coefficients of the polynomials that can be represented by a soft-plus unit. For example, Younes has shown that a soft-plus unit can represent a polynomial with an arbitrary leading coefficient:

Proposition 1 (Lemma 1 in [8]). *Let $B \subseteq V$ and $w_i = 0$ for $i \notin B$. Then, for any $J_B \in \mathbb{R}$, there is a choice of $w_B \in \mathbb{R}^B$ and $c \in \mathbb{R}$ such that $K_B = J_B$.*

Our goal is to show that we can actually choose the parameters in such a way that we can freely model two of the highest degree coefficients.

Let us first discuss the restrictions on the maximal degree, meaning that for some $B \subseteq V$ we require $K_C = 0$ for all $C \not\subseteq B$. We call a pair (B, B') an *edge pair* or a *covering pair* when $B \supset B'$ and there is no set C with $B \supsetneq C \supsetneq B'$.

Proposition 2. *Let (B, B') be an edge pair with $B' = B \setminus \{m\}$. Fixing $w_{B'} \in \mathbb{R}^{B'}$, $c \in \mathbb{R}$ and $w_{V \setminus B} = 0 \in \mathbb{R}^{V \setminus B}$, the equation $K_B = 0$ is satisfied either for at most $|B'|$ or for all values of w_m. A trivial solution is $w_m = 0$.*

Proof. Observe that

$$K_B(w, c) = K_B(w_B, c) = K_{B'}(w_{B'}, c + w_m) - K_{B'}(w_{B'}, c).$$

Hence $K_B = 0$ if and only if $K_{B'}(w_{B'}, c + w_m) = K_{B'}(w_{B'}, c)$. This has a trivial solution $w_m = 0$. To prove the upper bound on the number of solutions, let us write $K_{B'}(w_{B'}, c) = r$. We have

$$\begin{aligned} K_{B'}(w_{B'}, c + w_m) &= \sum_{C \subseteq B'} (-1)^{|B' \setminus C|} \log(1 + \exp(\sum_{i \in C} w_i + c + w_m)) \\ &= \log(\prod_{C \subseteq B'} (1 + \tilde{w}_m \tilde{c} \prod_{i \in C} \tilde{w}_i)^{(-1)^{|B' \setminus C|}}). \end{aligned}$$

Here we use the abbreviation $\tilde{r} = e^r$. Keep in mind that this is always positive. Now, $K_{B'}(w_{B'}, c + w_m) = r$ if and only if

$$\prod_{C \subseteq B'} (1 + \tilde{w_m} \tilde{c} \prod_{i \in C} \tilde{w}_i)^{(-1)^{|B' \setminus C|}} = \tilde{r},$$

or, equivalently,

$$\prod_{\substack{C \subseteq B': \\ B' \setminus C \text{ even}}} (1 + \tilde{w}_m \tilde{c} \prod_{i \in C} \tilde{w}_i) \quad - \quad \tilde{r} \prod_{\substack{C \subseteq B': \\ B' \setminus C \text{ odd}}} (1 + \tilde{w}_m \tilde{c} \prod_{i \in C} \tilde{w}_i) = 0.$$

This is a polynomial of degree at most $|B'|$ in $\tilde{w}_m$. □

The idea of Younes' proof of Proposition 1 is to choose all non-zero w_i of equal magnitude. In order to simplify the Möbius inversion formula, we choose the parameters w and c in such a way that the function ϕ has many zeros. Clearly this can only be done in an approximate way, since the soft-plus function is strictly positive. Nevertheless, these approximations can be made arbitrarily accurate, as $\log(1 + \exp(s)) \leq \exp(s)$ is arbitrarily close to zero for sufficiently large negative values of s.

The next lemma shows that the two highest degree coefficients can be modeled jointly by a soft-plus unit, at least in part. When the maximum degree $|B|$ is at most 3, the two coefficients are restricted by an inequality, but when $|B| \geq 4$, there are no such restrictions. The result is illustrated in Figure 2.

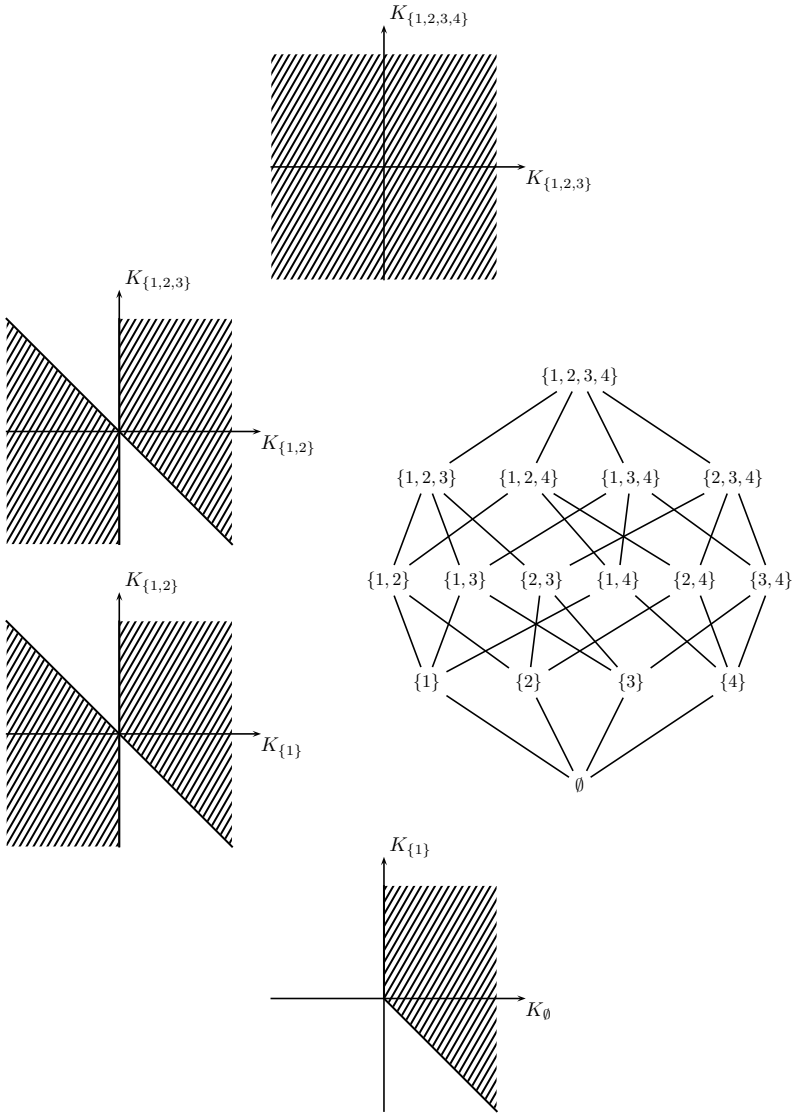


Figure 2: Illustration of Lemma 3. Depicted is for each edge pair (B, B') the set of all $(K_B, K_{B'}) \in \mathbb{R}^2$ for which there are some $K_C \in \mathbb{R}$, $C \neq B, B'$, such that the polynomial $\sum_{C \subseteq B} K_C \prod_{i \in C} x_i$ can be approximated arbitrarily well by a function of the form $\log(1 + \exp(\sum_{i \in B} w_i x_i + c))$.

Lemma 3. *Consider an edge pair (B, B'). Let $w_i = 0$ for $i \notin B$. Then, depending on $|B'|$, for any $\epsilon > 0$ there is a choice of w_B and c such that $\|(K_B, K_{B'}) - (J_B, J_{B'})\| \leq \epsilon$ if and only if*

$$\begin{aligned}
J_{B'} \geq 0 \ \wedge \ J_B \geq -J_{B'}, & \qquad \text{for } |B'| = 0 \\
J_{B'} \geq 0 \ \wedge \ J_B \geq -J_{B'} \quad \text{or} \quad J_{B'} \leq 0 \ \wedge \ J_B \leq -J_{B'}, & \qquad \text{for } |B'| = 1 \\
J_{B'} \geq 0 \ \wedge \ J_B \geq -J_{B'} \quad \text{or} \quad J_{B'} \leq 0 \ \wedge \ J_B \leq -J_{B'}, & \qquad \text{for } |B'| = 2 \\
(J_B, J_{B'}) \in \mathbb{R}^2, & \qquad \text{for } |B'| \geq 3.
\end{aligned}$$

Proof. Let $B' = B \setminus \{m\}$. The realizable edge coefficients satisfy

$$K_{B'}(w_{B'}, c) = \sum_{C \subseteq B'} (-1)^{|B' \setminus C|} \log(1 + \exp(\sum_{i \in C} w_i + c))$$

and

$$K_B(w_B, c) = K_{B'}(w_{B'}, c + w_m) - K_{B'}(w_{B'}, c).$$

Using this structure, we now proceed with the proof of the individual cases.

The case $|B'| = 0$. We omit this simple exercise.

The case $|B'| = 1$. The *if* statement is as follows. The elements of the set $\{0,1\}^B$ are the vertices of the $|B|$-dimensional unit cube. We call two vectors $x, x' \in \{0,1\}^B$ adjacent if they differ in exactly one entry, in which case they are the vertices of an edge of the cube.

The weights w_B and c can be chosen such that the affine map $\{0,1\}^B \to \mathbb{R};\ x_B \mapsto w_B^\top x_B + c$ maps two adjacent vectors to any arbitrary values and all other vectors to large negative values. The soft-plus function is monotonically increasing, taking value zero at minus infinity and plus infinity at plus infinity. Hence, for any $s, s' \in \mathbb{R}_+$, one finds weights w and c such that

$$\phi(x) = \begin{cases} s, & (x_{B'}, x_m) = (1, \ldots, 1, 1) \\ s', & (x_{B'}, x_m) = (1, \ldots, 1, 0) \\ \approx 0, & \text{otherwise} \end{cases},$$

or, alternatively, such that

$$\phi(x) = \begin{cases} s, & (x_{B'}, x_m) = (1, \ldots, 1, 0, 1) \\ s', & (x_{B'}, x_m) = (1, \ldots, 1, 0, 0) \\ \approx 0, & \text{otherwise} \end{cases}.$$

This leads to $K_B \approx (s - s')$ and $K_{B'} \approx s'$ or, alternatively, $K_B \approx -(s - s')$ and $K_{B'} \approx -s'$. The approximation can be made arbitrarily precise.

The *only if* statement is as follows. Denote the soft-plus function by $f \colon \mathbb{R} \to \mathbb{R}_+;\ s \mapsto \log(1 + \exp(s))$. We have that $K_{B'}(w_{B'}, c) = f(w_{B'} + c) - f(c)$ and $K_{B'}(w_{B'}, c + w_m) = f(w_{B'} + c + w_m) - f(c + w_m)$ are either both positive or both negative, depending on the sign of $w_{B'}$. If both are positive, then $K_B(w_B, c) = K_{B'}(w_{B'}, c + w_m) - K_{B'}(w_{B'}, c) \geq -K_{B'}(w_{B'}, c)$, and similarly in the case that both are negative.

The case $|B'| = 2$. The *if* statement follows from the previous case $|B'| = 1$. Indeed, consider an edge pair (C, C') with an element more than the edge pair (B, B'), such that $B = C\backslash\{n\}$ and $B' = C'\backslash\{n\}$. Then, for any w_B and c, choosing w_n large enough one obtains an arbitrarily accurate approximation $K_C((w_B, w_n), c - w_n) \approx K_B(w_B, c)$ and $K_{C'}((w_{B'}, w_n), c - w_n) \approx K_{B'}(w_{B'}, c)$.

For the *only if* statement we use a similar argument as previously. We have $K_{B'}(w_{B'}, c) = f(w_1 + w_2 + c) + f(c) - f(c + w_1) - f(c + w_2)$. By convexity of f, this is non-negative if and only if either $w_1, w_2 \geq 0$ or $w_1, w_2 \leq 0$. In other words, this is non-negative if and only if $w_1 \cdot w_2 \geq 0$. Under either of these conditions, $K_{B'}(w_{B'}, c + w_m)$ is also non-negative. Similarly, $K_{B'}(w_{B'}, c)$ is non-positive if and only if $w_1 \cdot w_2 \leq 0$. In this case, $K_{B'}(w_{B'}, c + w_m)$ is also non-positive. Now the statement follows as in the case $|B'| = 1$.

The case $|B'| \geq 3$. We need to show that all edge pairs are representable. Consider first $J_{B'} \geq 0$. We choose weights of the form $w_{B'} = \omega \mathbb{1}_{B'}$. Then $K_{B'}(w_{B'}, c) = f(3\omega + c) - 3f(2\omega + c) + 3f(\omega + c) - f(c)$. We can choose ω and c such that $3\omega + c = f^{-1}(J_{B'})$ while $2\omega + c, \omega c, c$ take very large negative values. This yields $K_{B'} \approx J_{B'}$.

Note that the derivative of the soft-plus function is $f'(s) = 1/(1 + \exp(-s))$, the logistic function. Choosing ω large enough from the beginning, the function $w_m \mapsto K_{B'}(w_{B'}, c + w_m)$ is monotonically increasing in the interval $w_m \in [0, \omega/2]$ and surpasses the value $\frac{1}{5}\omega$. On the other hand, when w_m is large enough, depending on ω and c, we have that $2\omega + c + w_m \geq \frac{5}{12}(3\omega + c + w_m)$ and $f(2\omega + c + w_m) \geq \frac{5}{12} f(3\omega + c + w_m)$. In this case $f(3\omega + c + w_m) - 3f(2\omega + c + w_m) \leq -\frac{1}{4}(3\omega + c + w_m) \leq -\frac{1}{4}\omega$. At the same time, $\omega + c + w_m$ and $c + w_m$ are smaller than $-\frac{1}{12}\omega$ and so $f(\omega + c + w_m)$ and $f(c + w_m)$ are very small in absolute value.

By the mean value theorem, depending on w_m, $K_{B'}(w_{B'}, c + w_m)$ takes any value in the interval $[-\frac{1}{5}\omega, \frac{1}{5}\omega]$, where ω is arbitrarily large. In turn, we can obtain $K_B = K_{B'}(w'_B, c + w_m) - K_{B'}(w_{B'}, c) \approx J_B$ for any $J_B \in \mathbb{R}$.

For $J_{B'} \leq 0$ the proof is analogous after label switching for one variable. □

It is also possible to control two maximal coefficients of the same degree:

Proposition 4. *Let $B, B' \subset V$ with $|B| = |B'| = 2$ and $|B \cup B'| = 3$. Let $w_i = 0$ for $i \notin B \cup B'$. Then for any $(J_B, J_{B'}) \in \mathbb{R}^2$ and $\epsilon > 0$ there is a choice of $w_{B \cup B'}$ and c such that $\|(K_B, K_{B'}) - (J_B, J_{B'})\| \leq \epsilon$ and $|K_C| \leq \epsilon$ for $C \notin B$ and for $C \notin B'$.*

Proof. Denote the soft-plus function by $f\colon \mathbb{R} \to \mathbb{R}_+;\ s \mapsto \log(1 + \exp(s))$. We will use the facts that $f(s) \approx 0$ when $s \ll -1$ and $f(s) \approx s$ when $s \gg 1$. In fact, note that $f(s) \leq \exp(s)$ and $f(s) - s = \log(1 + \exp(s)) - \log(\exp(s)) \leq \exp(-s)$.

Without loss of generality let $B = \{1, 2\}$ and $B' = \{2, 3\}$. Consider weights $w_1 = J_{\{1,2\}}$, $w_2 = 2\omega$, $w_3 = J_{\{1,3\}}$, and $c = -\omega$, for some ω. Then

$$
\begin{aligned}
&K_{\{1,2,3\}} \\
&= f(w_1 + w_2 + w_3 + c) - f(w_1 + w_2 + c) - f(w_2 + w_3 + c) - f(w_1 + w_3 + c) \\
&\quad + f(w_1 + c) + f(w_2 + c) + f(w_3 + c) - f(c)
\end{aligned}
$$

$$= f(J_{\{1,2\}} + J_{\{1,3\}} + \omega) - f(J_{\{1,2\}} + \omega) - f(J_{\{1,3\}} + \omega)$$
$$- f(J_{\{1,2\}} + J_{\{1,3\}} - \omega) + f(J_{\{1,2\}} - \omega) + f(\omega) + f(J_{\{1,3\}} - \omega) - f(-\omega).$$

Choosing $\omega \gg |J_{\{1,2\}}| + |J_{\{1,3\}}|$ we get

$$K_{\{1,2,3\}} \approx (J_{\{1,2\}} + J_{\{1,3\}} + \omega) - (J_{\{1,2\}} + \omega) - (J_{\{1,3\}} + \omega) + (\omega) = 0.$$

Similarly we get

$$\begin{aligned} K_{\{1,3\}} =& f(w_1 + w_3 + c) - f(w_1 + c) - f(w_3 + c) - f(c) \\ =& f(J_{\{1,2\}} + J_{\{1,3\}} - \omega) - f(J_{\{1,2\}} - \omega) - f(J_{\{1,3\}} - \omega) + f(-\omega) \approx 0 \end{aligned}$$

On the other hand,

$$\begin{aligned} K_{\{1,2\}} =& f(w_1 + w_2 + c) - f(w_1 + c) - f(w_2 + c) + f(c) \\ =& f(J_{\{1,2\}} + \omega) - f(J_{\{1,2\}} - \omega) - f(\omega) + f(-\omega) \approx J_{\{1,2\}}. \end{aligned}$$

Similarly, $K_{\{2,3\}} \approx J_{\{2,3\}}$. □

The intuition behind Proposition 4 is fairly simple. Consider the model with three binary visible variables, each interacting pairwise with the same hidden binary variable. This is the set of distributions of the form

$$p(x_1, x_2, x_3) = \sum_{y \in \{0,1\}} q(x_1|y) r(x_2|y) s(x_3|y) t(y).$$

Fixing $r(x_2|y) = \delta_{x_2,y}$, one obtains the set of distributions of the form

$$p(x_1, x_2, x_3) = q(x_1|x_2) s(x_3|x_2) t(x_2),$$

which correspond to the hierarchical model of three binary visible variables with pairwise interactions between the second and the first and between the second and the third.

It is natural to ask whether it is also possible to control other pairs of coefficients $K_B, K_{B'}$ of the same degree $|B| = |B'|$. In another direction, we would like to control triples of coefficients. In the analysis presented above, we ignore many of the degrees of freedom by moving many values of the soft-plus unit to zero. On the other hand, our analysis shows that, if $|B| = 3$ and $w_i = 0$ for $i \notin B$, then, despite having $|B| + 1 = 4$ parameters $w_i, i \in B$ and c to vary, we can only determine the two largest polynomial coefficients up to a certain inequality. We expect that the same is true in general: If we want to freely control k polynomial coefficients, we need strictly more than k parameters. Otherwise, the possible tuples of polynomial coefficients are restricted by some inequalities. The situation is well known in mixture models, which may require many more parameters to eliminate the corresponding inequalities than would be expected from naïve parameter counting [3].

4 Conditionally Independent Hidden Variables

In the case of a bipartite graph between V and H with all variables binary, the hierarchical model (or its visible marginal) is called a restricted Boltzmann machine, denoted $\mathrm{RBM}_{V,H}$. The free energy takes the form

$$F(x) = \sum_{j\in H} \log\left(1 + \exp\Big(\sum_{i\in V} w_{ji}x_i + c_j\Big)\right) + \sum_{i\in V} b_i x_i.$$

This is the sum of an arbitrary degree-one polynomial, with coefficients b_i, $i \in V$ (biases of the visible variables), and H independent soft-plus units, with parameters w_{ji}, $j \in H, i \in V$ (coupling strengths), c_j, $j \in H$ (biases of the hidden variables). We can use each soft-plus unit to model a group of coefficients of a given polynomial, as explained in Section 3, starting at the highest degrees. In view of Lemma 3 and Proposition 4, the problem of representing a polynomial can be reduced to covering the appearing monomials by pairs of coefficients that can be jointly controlled. If we can find a disjoint covering, then it suffices to add $H = \frac{1}{2}|\{C \in S : |C| \geq 2\}|$ hidden variables. However, it may not always be possible to choose a disjoint covering. So in general, we are led to the following technical theorem:

Theorem 5. *Consider a hierarchical model $\mathcal{E}_S$ on $\{0,1\}^V$. Then every distribution from $\mathcal{E}_S$ can be approximated arbitrarily well by distributions from $\mathrm{RBM}_{V,H}$ whenever $|H| \geq N + M$, where N is the minimal number of pairs (B, B') with $B \supset B'$, $|B| = |B'| + 1$, $|B'| \geq 3$, that cover $\{C \in S\colon |C| \geq 3\}$ and M is minimal number of pairs (B, B') with $|B| = |B'| = 2$, $|B \cap B'| = 1$, that cover $\{C \in S\colon |C| = 2\}$.*

The problem of finding a minimal covering is combinatorial. For the k-interaction model, where $S = \{\Lambda \subseteq V\colon |\Lambda| \leq k\}$, we have the following upper bound:

Corollary 6. *Let $3 \leq k \leq |V|$. Then every distribution from the k-interaction model can be approximated arbitrarily well by distributions from $\mathrm{RBM}_{V,H}$ whenever*

$$|H| \geq \sum_{j=2}^{k} \binom{|V|-1}{j} + \frac{1}{2}\binom{|V|}{2}.$$

If $k = 2$, then $|H| \geq \frac{1}{2}\binom{|V|}{2}$ is sufficient.

Proof. The set 2^V of subsets of V can be identified with the set $\{0,1\}^V$ of their indicator functions. The set 2^V is partially ordered by inclusion. The corresponding Hasse diagram has the same edges as the binary cube $\{0,1\}^V$. The diagram has levels corresponding to the cardinality of its elements. Consider the set of edges of the form $((0, x_2, \ldots, x_V), (1, x_2, \ldots, x_V))$. At level j there are $\binom{|V|-1}{j}$ such edges going upwards and $\binom{|V|-1}{j-1}$ going downwards. Hence $\sum_{j=2}^{\min\{k,|V|-1\}} \binom{|V|}{j}$ edges cover all elements of cardinality $3 \leq |B| \leq k$. By Lemma 3, each of the corresponding coefficient pairs can be modeled with one hidden variable.

On the other hand, there are $\binom{|V|}{2}$ cardinality-two subsets of V. This set can be divided into $\lfloor \frac{1}{2}\binom{|V|}{2} \rfloor$ pairs of overlapping sets plus possibly one more set. By Proposition 4 each of the corresponding coefficient pairs, or an individual coefficient, can be modeled with one hidden variable. □

We can also consider models that include interactions among the visible variables other than just the biases. In this case we only need to cover the interaction sets from S that are not already included in T. In Theorem 5 one just replaces S by $S \setminus T$. We note the following special case:

Corollary 7. *Each distribution from the k-interaction model can be approximated arbitrarily well by distributions from a pairwise interaction model with $|H| = \sum_{j=2}^{k} \binom{|V|-1}{j}$ hidden binary variables.*

Proof. The arguments are exactly as in the proof of Corollary 6, except that here we consider an approximating model with full pairwise interactions among its visible variables. □

Remark 8. In general an RBM contains many more distributions than just the interaction models indicated in the corollary. For instance, an RBM with $|H| \geq K$ hidden variables can approximate any distribution with support of cardinality K arbitrarily well. On the other hand, every distribution with support of cardinality K is contained in the closure of the k-interaction model if and only if $2^k - 1 \geq K$, see [2]. Using the corollary we would need $|H| \geq \sum_{j=2}^{k} \binom{|V|-1}{j} + \frac{1}{2}\binom{|V|}{2}$ to represent this model. This can be much larger than $2^k - 1$ when $|V| - 1$ is larger than k.

We present a few examples illustrating our results.

Example 9 ($\mathrm{RBM}_{3,1}$)**.** The restricted Boltzmann machine with $|V| = 3$ visible variables and $|H| = 1$ hidden variables is the same as the 2-mixture of product distributions of three binary variables, which is also known as the *tripod tree model*. It has 7 parameters and the same dimension. What is the largest hierarchical model contained in this model?

It contains any hierarchical model with a single pairwise interaction. This can be explained from our results as follows. The degree-two coefficient can be modeled with one soft-plus unit (Proposition 1), whereas the linear coefficients can be modeled with the biases of the visible variables. An alternative way to see this is that the 2-mixture of product distributions of two binary variables is equal to the set of all joint distributions of two binary variables.

It contains each of the three hierarchical models with two pairwise interactions. Two degree-two coefficients with one shared variable can be jointly modeled by one soft-plus unit (Proposition 4), whereas the linear coefficients can be modeled with the biases of the visible variables.

It does not contain the hierarchical model with three pairwise interactions, which is known as the *no three way interaction model*. One way of proving this is by comparing the possible support sets of the two models, as proposed in [3]: The support set of

a mixture of two product distributions is a union of two cylinder sets. On the other hand, the possible support sets of a hierarchical model correspond to the faces of its marginal polytope. The marginal polytope of the no three way interaction model is the cyclic polytope $C(8,6)$, which has $N = 8$ vertices and dimension $d = 6$ (see, e.g., [3, Lemma 18]). This is a neighborly polytope, meaning that every $d/2 = 3$ or less vertices form a face, or that every subset of $\{0,1\}^3$ of cardinality $d/2 = 3$ is the support set of a distribution in the closure of the model.[1]The claim then follows from the fact that the set $\{(100),(010),(001)\}$ is not a union of two cylinder sets.

Example 10 ($\mathrm{RBM}_{3,2}$)**.** This model contains the no three way interaction model. Two of the quadratic coefficients can be jointly modeled by one soft-plus unit (Proposition 4). The remaining quadratic coefficient can be modeled by one soft-plus unit (Proposition 1). The linear coefficients can be modeled with the biases of the visible variables.

It does not contain the full interaction model. This can be deduced from analyzing the possible support sets of the distributions in the closure of the RBM model. For details on this interesting subject we refer the reader to [6].

Example 11 ($\mathrm{RBM}_{3,3}$)**.** This model is a universal approximator; see [4]. This observation can be recovered from our results as follows. Two degree-two coefficients can be jointly modeled with one soft-plus unit (Proposition 4). The degree-three and the remaining degree-two coefficients can each be modeled with one soft-plus unit (Proposition 1). Finally the linear coefficients can be modeled with the biases of the visible variables.

Example 12 ($\mathrm{RBM}_{4,7}$)**.** This model is a universal approximator; see [4]. Our results recover observation this as follows. The 6 quadratic coefficients can be grouped into 3 pairs with a shared variable in each pair. By Proposition 4 these can be modeled with 3 soft-plus units. By Lemma 3 the quartic and one cubic coefficients can be modeled with one soft-plus unit. By Proposition 1 the remaining 3 cubic coefficients can be modeled with one soft-plus unit each.

5 Conclusions

We have studied what kind of interactions can appear when marginalizing over a hidden variable that is connected by pair-interactions with all visible variables. We have focused on controlling two interactions at a time. The examples at the end of Section 4 show that our analysis gives tight results in many cases. These results generalize and improve the analysis from [4] and [8], respectively. They can also be easily extended to improve previous considerations for conditional probability distributions [5]. On the other hand, many questions are still open at this point, and a full characterization of soft-plus polynomials and the necessary number of hidden variables is missing. Many other questions are left open:

[1]More generally, in [2] it is shown that if $k+1$ is the smallest cardinality of a non-face of S, then the marginal polytope of $\mathcal{E}_S$ is $2^k - 1$ neighborly, meaning that any $2^k - 1$ or fewer of its vertices define a face.

It would be interesting to look at non-binary hidden variables. This corresponds to analyzing the hierarchical models that can be represented by mixture models. In the case of binary hidden variables, the partial factorization leads to soft-plus units, whereas in the case of larger hidden variables, it will lead straight to a shifted logarithm of denormalized mixtures. Similarly, it would be interesting to take a look at non-binary visible variables. In this case state vectors cannot be identified with subsets of units. This means that the correspondence between function values and polynomial coefficients is not as direct.

Some of the general considerations presented here can be applied to obtain simple results on the representation of hierarchical models in terms of hierarchical models with hidden variables and more than pairwise interactions, even though the case of pairwise interactions is the more interesting one from the perspective of distributed networks and efficient Gibbs sampling. Another interesting direction are models where the hidden variables are not conditionally independent given the visible variables, e.g. models involving several layers of hidden variables like the deep Boltzmann machines. This case is more challenging, since the free energy does not decompose into independent terms.

Acknowledgments

We thank Nihat Ay for helpful remarks with the manuscript.

References

[1] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721–741, 1984.

[2] T. Kahle. Neighborliness of marginal polytopes. *Beiträge zur Algebra und Geometrie*, 51(1):45–56, 2010.

[3] G. Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *Kybernetika*, 49(1):23–39, 2013.

[4] G. Montúfar and N. Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 23(5):1306–1319, 2011.

[5] G. Montúfar, N. Ay, and K. Ghazi-Zahedi. Geometry and expressive power of conditional restricted Boltzmann machines. *Journal of Machine Learning Research*, 2015. To appear. arXiv preprint arXiv:1402.3346.

[6] G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics*, 29:321–347, 2015.

[7] B. Steudel and N. Ay. Information-theoretic inference of common ancestors. *Entropy*, 17(4):2304, 2015.

[8] L. Younes. Synchronous Boltzmann machines can be universal approximators. *Applied Mathematics Letters*, 9(3):109 – 113, 1996.

[9] P. Zwiernik and J. Q. Smith. Tree cumulants and the geometry of binary tree models. *Bernoulli*, 18:290–321, 2012.

Mode Poset Probability Polytopes

Guido Montúfar
Max Planck Institute for
Mathematics in the Sciences
montufar@mis.mpg.de

Johannes Rauh
Department of Mathematics and Statistics
York University
jarauh@yorku.ca

Abstract

A mode of a probability vector is an elementary event that has more probability mass than each of its direct neighbors, with respect to some vicinity structure on the set of elementary events. The mode inequalities cut out a polytope from the simplex of probability vectors. Related to this is the concept of strong modes. A strong mode of a distribution is an elementary event that has more probability mass than all its direct neighbors together. The set of probability distributions with a given set of strong modes is again a polytope. We study the vertices, the facets, and the volume of such polytopes depending on the sets of (strong) modes and the vicinity structures.

1 Introduction

Many probability models used in practice are given in a parametric form. Sometimes it is useful to also have an implicit description in terms of properties that characterize the probability distributions that belong to the model. Such a description can be used to check whether a given probability distribution lies in the model or, otherwise, to estimate how far it lies from the model. For example, if a given model has a parametrization by polynomial functions, then one can show that it has a *semialgebraic description*; that is, an implicit description as the solution set of polynomial equations and polynomial inequalities. Finding this description is known as the *implicitization* problem, which in general is very hard to solve completely. Even if it is not possible to give a full implicit description, it may be possible to confine the model by simple polynomial equalities and inequalities. Here we are interested in simple confinements, in terms of natural classes of linear equalities and inequalities.

We consider polyhedral sets of discrete probability distributions defined by prescribed sets of modes. A mode is a local maximum of a probability vector. Locality is with respect to a given a vicinity structure in the set of coordinate indices; that is, x is a (strict) mode of a probability vector p if and only if $p_x > p_y$, for all neighbors y of x. The vicinity structure depends on the setting. For probability distributions on a set of fixed-length strings, it is natural to call two strings neighbors if and only if they have Hamming distance one. For probability distributions on integer intervals, it is

natural to call two integers neighbors if and only if they are consecutive. In general, a vicinity structure is just a graph with undirected edges.

Modes are important characteristics of probability distributions. In particular, the question whether a probability distribution underlying a statistical experiment has one or more modes is important in applications. Also, many statistical models consist of "nice" probability distributions that are "smooth" in some sense. Such probability distributions have only a limited number of modes. Another motivation for studying modes was given in [2], where it was observed that mode patterns are a practical way to differentiate between certain parametric model classes.

Besides from modes, we are also interested in the related concept of strong modes introduced in [2]. A point x is a (strict) strong mode of a probability distribution p if and only if $p_x > \sum_{y\sim x} p_y$, where the sum runs over all neighbors y of x. Strong modes offer similar possibilities as modes for studying models of probability distributions. While strong modes are more restrictive than modes, they are easier to study.

One observation is: Suppose that $p = \sum_{i=1}^{k} \lambda_i p^i$ is a mixture of k probability distributions. If p has a strict strong mode $x \in V$, then x must be a mode of one of the distributions p^i, because if $p^i(x) \leq p^i(y_i)$ for some neighbor y_i of x for all i, then $\sum_i \lambda_i p^i(x) \leq \sum_i \lambda_i p^i(y_i) \leq \sum_{y\sim x} \sum_i \lambda_i p^i(y)$. For example, a mixture of k uni-modal distributions has at most k strong modes. Surprisingly, the same statement is not true for modes: A mixture of k product distributions may have more than k modes [2]. Still, the number of modes of a mixture of product distributions is bounded, although this bound is not known in general. As another example, in [2] it was shown that a restricted Boltzmann machine with m hidden nodes and n visible nodes, where $m < n$ and m is even, does not contain probability distributions with certain patterns of 2^m strict strong modes.

In this paper we derive essential properties of (strong) mode polytopes, depending on the vicinity structures and the considered patterns of (strong) modes. In particular, we describe the vertices, the facets, and the volume of these polytopes. It is worth mentioning that mode probability polytopes are closely related to order and poset polytopes. We describe this relation at the end of Section 2.

This paper is organized as follows: In Section 2 we study the polytopes of modes and in Section 3 the polytopes of strong modes.

2 The Polytope of Modes

We consider a finite set of elementary events V and the set of probability distributions on this set, $\Delta(V)$. We endow V with a vicinity structure described by a graph. Let $G = (V, E)$ be a simple graph (i.e., no multiple edges and no loops). For any $x, y \in V$, if $(x, y) \in E$ is an edge in G, we write $x \sim y$. Since we assume that the graph is simple, $x \sim y$ implies $x \neq y$.

Definition 1. A point $x \in V$ is a *mode* of a probability distribution $p \in \Delta(V)$ if $p_x \geq p_y$ for all $y \sim x$.

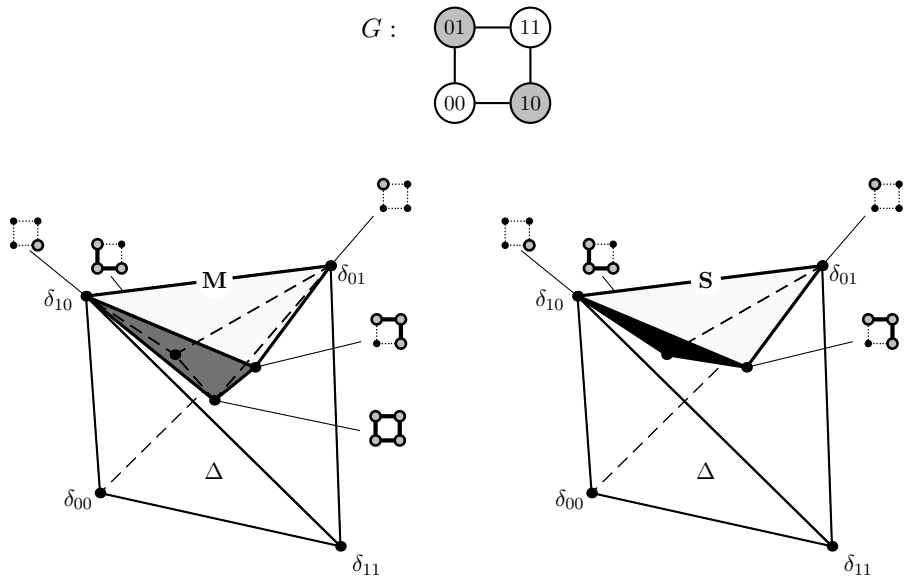


Figure 1: Above: The graph G from Examples 3 and 11, with $\mathcal{C}$ marked in gray. Below: The corresponding polytopes $\mathbf{M}(G,\mathcal{C})$ and $\mathbf{S}(G,\mathcal{C})$. Each vertex of these polytopes is a uniform distribution supported on a subset of G, as explained in Propositions 4 and 12.

Definition 2. Consider a subset $\mathcal{C} \subseteq V$. The *polytope of $\mathcal{C}$-modes* in G is the set $\mathbf{M}(G,\mathcal{C})$ of all probability distributions $p \in \Delta(V)$ for which every $x \in \mathcal{C}$ is a mode.

The set $\mathbf{M}(G,\mathcal{C})$ is always non-empty, since it contains the uniform distribution. It is a polytope, because it is a closed convex set defined by finitely many linear inequalities and, as a subset of $\Delta(V)$, it is bounded. We are interested in the properties of this polytope, depending on G and $\mathcal{C}$.

Recall that a set of vertices of a graph is *independent*, if it does not contain two adjacent elements. If $\mathcal{C}$ is not independent, then $\mathbf{M}(G,\mathcal{C})$ is not full-dimensional as a subset of $\Delta(V)$; that is, $\dim \mathbf{M}(G,\mathcal{C}) < \dim(\Delta(V)) = |V| - 1$. For, if $x, y \in \mathcal{C}$ are neighbors, then the defining equations of $\mathbf{M}(G,\mathcal{C})$ imply that $p_x \geq p_y \geq p_x$; that is, any $p \in \mathbf{M}(G,\mathcal{C})$ satisfies $p_x = p_y$. In the following we will ignore this degenerate case and assume that the set of modes is independent.

In some applications, for example those mentioned in the introduction, it is more natural to study *strict modes*; i.e. points $x \in V$ with $p_x > p_y$ for all $y \sim x$. A description of the set of distributions with prescribed strict modes is easy to obtain from a description of $\mathbf{M}(G,\mathcal{C})$.

Example 3. Let G be a square with vertices $V = \{00, 01, 10, 11\}$ and edges $E = \{(00,01), (00,10), (01,11), (10,11)\}$. The polytope $\mathbf{M}(G,\mathcal{C})$ for $\mathcal{C} = \{01, 10\}$ is given in Figure 1.

Vertices

We have defined $\mathbf{M}(G,\mathcal{C})$ by linear inequalities (H-representation). Next we determine its vertices (V-representation). For any non-empty $W \subseteq V \setminus \mathcal{C}$ and $y \in V$ write $y \sim W$ if $y \sim x$ for some $x \in W$. Moreover, let $N_{\mathcal{C}}(W) = \{y \in \mathcal{C} : y \sim W\}$ (this is the set of declared modes which are neighbors of W), and let $e_{\mathcal{C}}^{W}$ be the uniform distribution on $N_{\mathcal{C}}(W) \cup W$.

Proposition 4.

1. $\mathbf{M}(G,\mathcal{C})$ *is the convex hull of* $\{e_{\mathcal{C}}^{W} : \emptyset \neq W \subseteq V \setminus \mathcal{C}\} \cup \{\delta_x : x \in \mathcal{C}\}$, *where* δ_x *denotes the point distribution concentrated on* x.

2. *For any* $x \in \mathcal{C}$, *the distribution* δ_x *is a vertex of* $\mathbf{M}(G,\mathcal{C})$.

3. $e_{\mathcal{C}}^{W}$ *is a vertex of* $\mathbf{M}(G,\mathcal{C})$ *iff for any* $x, y \in W$, $x \neq y$, *there is a path* $x = x_0 \sim x_1 \sim \cdots \sim x_r = y$ *in* G *with* $x_0, x_2, \cdots \in W$ *and* $x_1, x_3, \cdots \in N_{\mathcal{C}}(W)$.

Proof. Clearly, for every non-empty $W \subseteq V \setminus \mathcal{C}$, the vector $e_{\mathcal{C}}^{W}$ belongs to $\mathbf{M}(G,\mathcal{C})$, and the same is true for the vectors δ_x with $x \in \mathcal{C}$ ($\mathcal{C}$ is independent). Next we show that each $p \in \mathbf{M}(G,\mathcal{C})$ can be written as a convex combination of $\{e_{\mathcal{C}}^{W} : \emptyset \neq W \subseteq V \setminus \mathcal{C}\} \cup \{\delta_x : x \in \mathcal{C}\}$. We do induction on the cardinality of $W := \mathrm{supp}(p) \setminus \mathcal{C}$. If $|W| = 0$, then $p \in \Delta(\mathcal{C})$ is a convex combination of $\{\delta_x : x \in \mathcal{C}\}$. Now assume $|W| > 0$. Let $\lambda = \min\{p_x : x \in W\}$. Then, $p - \lambda e_{\mathcal{C}}^{W} \geq 0$ (component-wise) and $\sum_x (p_x - \lambda e_{\mathcal{C}}^{W}(x)) = (1 - \lambda)$. Therefore,

$$p' := \frac{1}{1-\lambda}(p - \lambda e_{\mathcal{C}}^{W}) \in \Delta(V).$$

Moreover, one checks that $p' \in \mathbf{M}(G,\mathcal{C})$. By definition, $\mathrm{supp}(p') \setminus \mathcal{C} \subsetneq \mathrm{supp}(p) \setminus \mathcal{C}$. By induction, $\mathrm{supp}(p')$ is a convex combination of $\{e_{\mathcal{C}}^{W} : \emptyset \neq W \subseteq V \setminus \mathcal{C}\} \cup \{\delta_x : x \in \mathcal{C}\}$, and so the same is true for p.

It remains to check which elements of $\{e_{\mathcal{C}}^{W} : \emptyset \neq W \subseteq V \setminus \mathcal{C}\} \cup \{\delta_x : x \in \mathcal{C}\}$ are vertices of $\mathbf{M}(G,\mathcal{C})$. Since δ_x is a vertex of $\Delta(V)$, it is also a vertex of $\mathbf{M}(G,\mathcal{C})$. Let $W \subset V \setminus \mathcal{C}$ be non-empty. Call a path such as in the statement of the proposition an *alternating path*. Suppose that there is no alternating path from x to y for some $x, y \in W$. Let $W_1 = \{z \in W :$ there is an alternating path from x to $z\}$ and let $W_2 = W \setminus W_1$. Then W_1, W_2 are non-empty, and $N_{\mathcal{C}}(W_1) \cap \tilde{N}_{\mathcal{C}}(W_2)$ is empty. Hence $e_{\mathcal{C}}^{W}$ is a convex combination of $e_{\mathcal{C}}^{W_1}$ and $e_{\mathcal{C}}^{W_2}$, and $e_{\mathcal{C}}^{W}$ is not a vertex.

Let W be a non-empty subset of $V \setminus \mathcal{C}$ such that any pair of elements of W is connected by an alternating path. To show that $e_{\mathcal{C}}^{W}$ is a vertex, for any different non-empty set $W' \subseteq V \setminus \mathcal{C}$ we need to find a face of $\mathbf{M}(G,\mathcal{C})$ that contains $e_{\mathcal{C}}^{W}$ but not $e_{\mathcal{C}}^{W'}$. If there exists $x \in W' \setminus W$, then $e_{\mathcal{C}}^{W'}(x) > 0 = e_{\mathcal{C}}^{W}(x)$. Hence, $e_{\mathcal{C}}^{W}$ lies on the face of $\mathbf{M}(G,\mathcal{C})$ defined by $p_x \geq 0$, but $e_{\mathcal{C}}^{W'}$ does not. Otherwise, $W' \subsetneq W$. Let $x' \in W \setminus W'$ and $y' \in W' \neq \emptyset$. By assumption, there exists an alternating path from x' to y' in W. On this path, there exist $x \in W \setminus W'$ and $y \in \mathcal{C}$ with $y \sim x$ and $y \in N_{\mathcal{C}}(W')$. Therefore, $e_{\mathcal{C}}^{W'}(y) - e_{\mathcal{C}}^{W'}(x) > 0 = e_{\mathcal{C}}^{W}(y) - e_{\mathcal{C}}^{W}(x)$. □

Corollary 5. $\mathbf{M}(G,\mathcal{C})$ *is a full-dimensional sub-polytope of* $\Delta(V)$.

Proof. The convex hull of $\{\delta_x : x \in \mathcal{C}\} \cup \{e_{\mathcal{C}}^{\{y\}} : y \in V \setminus \mathcal{C}\}$ is a $(|V|-1)$-simplex and a subset of $\mathbf{M}(G,\mathcal{C})$. □

Facets

$\mathbf{M}(G,\mathcal{C})$ is defined, as a subset of $\Delta(V)$, by the inequalities

$p_x \geq 0$,	for all $x \in V$,	(positivity inequalities)
$p_x \geq p_y$,	for all $x \in \mathcal{C}$ and $y \sim x$.	(mode inequalities)

Next we discuss, which of these inequalities define facets.

Proposition 6.

1. *For any* $x \in V \setminus \mathcal{C}$, *the positivity inequality* $p_x \geq 0$ *defines a facet.*
2. *If* $x \in \mathcal{C}$, *then* $p_x \geq 0$ *defines a facet iff* x *is isolated in* G.
3. *For any* $x \in \mathcal{C}$ *and* $y \sim x$, *the mode inequality* $p_x \geq p_y$ *defines a facet.*

Proof. 1. For $x \in V \setminus \mathcal{C}$, the inequality $p_x \geq 0$ defines a facet of the subsimplex from the proof of Corollary 5, and hence also of $\mathbf{M}(G,\mathcal{C})$.

2. If x is isolated, then x is a mode of any distribution. Therefore, $\mathbf{M}(G,\mathcal{C}) = \mathbf{M}(\mathcal{C} \setminus \{x\})$, and the statement follows from 1.

Otherwise, suppose there exists $y \in V$ with $x \sim y$. Since $\mathcal{C}$ is independent, $y \notin \mathcal{C}$. Then $p_x = (p_x - p_y) + p_y$; that is, the inequality $p_x \geq 0$ is implied by the inequalities $p_x \geq p_y$ and $p_y \geq 0$, and $p_x \geq 0$ defines a sub-face of the facet $p_y \geq 0$, which is a strict sub-face, since it does not contain δ_x. Therefore, $p_x \geq 0$ does not define a facet itself.

3. Let $W := \{z \in \mathcal{C} : z \sim y\} \setminus \{x\}$. The uniform distribution on $W \cup \{y\}$ satisfies all defining inequalities of $\mathbf{M}(G,\mathcal{C})$, except $p_x \geq p_y$. □

Triangulation and volume

The polytope $\mathbf{M}(G,\mathcal{C})$ has a natural triangulation that comes from a natural triangulation of $\Delta(V)$. Let $N = |V|$ be the cardinality of V. For any bijection $\sigma : \{1,\dots,N\} \to V$ let

$$\Delta_\sigma = \{p \in \Delta(V) : p_{\sigma(i)} \leq p_{\sigma(i+1)} \text{ for } i = 1,\dots,N-1\}.$$

Clearly, the Δ_σ form a triangulation of $\Delta(V)$. In particular, $\Delta(V) = \bigcup_\sigma \Delta_\sigma$ and $\operatorname{vol}(\Delta_\sigma \cup \Delta_{\sigma'}) = \operatorname{vol}(\Delta_\sigma) + \operatorname{vol}(\Delta_{\sigma'})$ whenever $\sigma \neq \sigma'$.

Lemma 7. *Let* $\Sigma(G,\mathcal{C})$ *be the set of all bijections* $\sigma : \{1,\dots,N\} \to V$ *that satisfy* $\sigma^{-1}(x) < \sigma^{-1}(y)$ *for all* $y \in \mathcal{C}$ *and* $x \sim y$. *Then* $\mathbf{M}(G,\mathcal{C}) = \bigcup_{\sigma \in \Sigma(G,\mathcal{C})} \Delta_\sigma$.

Proof. If $\sigma \in \Sigma$ and $p \in \Delta_\sigma$, then $p \in \mathbf{M}(G, \mathcal{C})$ by definition. Conversely, let $p \in \mathbf{M}(G, \mathcal{C})$. Choose a bijection $\sigma : \{1, \ldots, N\} \to V$ that satisfies the following:

1. $p_{\sigma(i+1)} \geq p_{\sigma(i)}$ for $i = 1, \ldots, N-1$,
2. If $x \in \mathcal{C}$ and $y \sim x$, then $\sigma^{-1}(x) \leq \sigma^{-1}(y)$.

Clearly, $\sigma \in \Sigma$, and $p \in \Delta_\sigma$. □

Corollary 8. $\operatorname{vol}(\mathbf{M}(G, \mathcal{C})) = \frac{|\Sigma|}{|V|!} \operatorname{vol}(\Delta(V))$.

Proof. All simplices Δ_σ have the same volume. Moreover, $\operatorname{vol}(\Delta_\sigma \cap \Delta_{\sigma'}) = 0$ for $\sigma \neq \sigma'$. Thus, $\operatorname{vol}(\mathbf{M}(G, \mathcal{C})) = |\Sigma| \operatorname{vol}(\Delta_\sigma)$ and $\operatorname{vol}(\Delta(V)) = |V|! \operatorname{vol}(\Delta_\sigma)$. □

It remains to compute the cardinality of $\Sigma(G, \mathcal{C})$. It is not difficult to enumerate $\Sigma(G, \mathcal{C})$ by iterating over the set V. However, $\Sigma(G, \mathcal{C})$ may be a very large, and so, enumerating it can take a very long time. In fact, this is a special instance of the problem of counting the number of linear extensions of a partial order (see below); a problem which in many cases is known to be $\#P$-complete [1]. In our case, a simple lower bound is $|\Sigma(G, \mathcal{C})| \geq |\mathcal{C}|! |V \setminus \mathcal{C}|!$ (equality holds only when G is a complete bipartite graph and $\mathcal{C}$ is one of the maximal independent sets).

Relation to order polytopes

The results in this section can also be derived from results about order polytopes. To explain this, it is convenient to slightly generalize our settings. Instead of looking at a graph G and an independent subset $\mathcal{C}$ of nodes, consider a partial order $\succeq$ on V and let

$$\mathbf{M}(\succeq) := \{p \in \Delta(V) : p_x \geq p_y \text{ whenever } x \succeq y\}.$$

The polytope $\mathbf{M}(G, \mathcal{C})$ arises in the special case where $\succeq$ is defined by

$$x \succeq y \quad :\Longleftrightarrow \quad x \sim y \text{ and } x \in \mathcal{C}.$$

The relation $\succeq$ defined in this way from G and $\mathcal{C}$ is a partial order precisely if $\mathcal{C}$ is independent. Our results about vertices, facets and volumes directly generalize to $\mathbf{M}(\succeq)$. We omit further details at this point.

The *order polytope* of a partial order arises by looking at subsets of the unit hypercube instead of subsets of the probability simplex (see [3] and references):

$$\mathcal{O}(\succeq) := \{p \in [0,1]^V : p_x \geq p_y \text{ whenever } x \succeq y\}.$$

One can show that $\mathbf{M}(\succeq)$ is the vertex figure of $\mathcal{O}(\succeq)$ at the vertex 0. This observation allows to transfer the results from [3] to $\mathbf{M}(G, \mathcal{C})$.

3 The Polytope of Strong Modes

Definition 9. A point $x \in V$ is a *strong mode* of a probability distribution $p \in \Delta(V)$ if $p_x \geq \sum_{y \sim x} p_y$.

Definition 10. Consider a subset $\mathcal{C} \subseteq V$. The *polytope of strong $\mathcal{C}$-modes* in G is the set $\mathbf{S}(G, \mathcal{C})$ all probability distributions $p \in \Delta(V)$ for which every $x \in \mathcal{C}$ is a strong mode.

Again, in applications one may be interested in *strict strong modes* that are characterized by strict inequalities of the form $p_x > \sum_{y \sim x} p_y$.

If $x \sim y$ for two strong modes of $p \in \Delta(V)$, then $p_x = p_y$ and $p_z = 0$ for all other neighbors z of x or y. In order to avoid such pathological cases, in the following we always assume that $\mathcal{C}$ is an independent subset of G.

Example 11. Consider the graph from Example 3. For $\mathcal{C} = \{01, 10\}$, the polytope $\mathbf{S}(G, \mathcal{C})$ is given in Figure 1.

Again, we are interested in the vertices of the polytope $\mathbf{S}(G, \mathcal{C})$. For any $x \in V$ let $N_{\mathcal{C}}(x) = \{y \in \mathcal{C} : y \sim x\}$ (this is the set of strong modes which are neighbors of x) and let $f^x_{\mathcal{C}}$ be the uniform distribution on $N_{\mathcal{C}}(x) \cup \{x\}$.

Proposition 12. *If $\mathcal{C}$ is independent, then $\mathbf{S}(G, \mathcal{C})$ is a $(|V|-1)$-simplex with vertices $f^x_{\mathcal{C}}$, $x \in V$.*

Proof. To see that $\{f^x_{\mathcal{C}} : x \in V\}$ is linearly independent, observe that the matrix with columns $f^x_{\mathcal{C}}$ is in tridiagonal form when V is ordered such that the vertices in $\mathcal{C}$ come before the vertices in $V \setminus \mathcal{C}$. Therefore, the probability distributions $f^x_{\mathcal{C}}$ span a $(|V|-1)$-dimensional simplex.

It is easy to check that $f^x_{\mathcal{C}} \in \mathbf{S}(G, \mathcal{C})$ for any $x \in V$. It remains to prove that any $p \in \mathbf{S}(G, \mathcal{C})$ lies in the convex hull of $\{f^x_{\mathcal{C}} : x \in V\}$. We do induction on the cardinality of $W := \operatorname{supp}(p) \setminus \mathcal{C}$. If $|W| = 0$, then $p \in \Delta(\mathcal{C})$ is a convex combination of $\{\delta_x : x \in \mathcal{C}\} = \{f^x_{\mathcal{C}} : x \in \mathcal{C}\}$. Otherwise, let $x \in W$. Then

$$p' := \frac{1}{1 - p_x}(p - p_x f^x_{\mathcal{C}}) \in \Delta(V),$$

since $p \in \mathbf{M}(G, \mathcal{C})$. Moreover, $p' \in \mathbf{M}(G, \mathcal{C})$. The statement now follows by induction, since $\operatorname{supp}(p') \setminus \mathcal{C} = W \setminus \{x\}$. □

Proposition 13. *The facets of $\mathbf{S}(G, \mathcal{C})$ are $p_x \geq \sum_{y \sim x} p_y$ for all $x \in \mathcal{C}$ and $p_x \geq 0$ for all $x \in V \setminus \mathcal{C}$.*

Proof. It is easy to verify that each of the faces defined by these inequalities contains $|V| - 1$ vertices. □

Proposition 14. $\operatorname{vol}(\mathbf{S}(G, \mathcal{C})) = \Big(\prod_{x \in V} \frac{1}{|N_{\mathcal{C}}(x)| + 1} \Big) \operatorname{vol}(\Delta(V))$.

Proof. After rearrangement of columns, the matrix

$$(f_{\mathcal{C}}^x)_{x\in V} = \left((\delta_x)_{x\in\mathcal{C}}, \left(\tfrac{1}{|N_{\mathcal{C}}(x)|+1}\mathbb{1}_{N_{\mathcal{C}}(x)}\right)_{x\in V\setminus\mathcal{C}, x\sim\mathcal{C}}, (\delta_x)_{x\in V\setminus\mathcal{C}, x\not\sim\mathcal{C}} \right)$$

is in upper triangular from, with diagonal elements $\frac{1}{|N_{\mathcal{C}}(x)|+1}$, $x\in V$. The statement now follows from the next Lemma 15. □

Lemma 15. *Let $\Delta = \operatorname{conv}\{e_0,\dots,e_d\}$ be the standard d-simplex in $\mathbb{R}^{d+1}$ and let $s_0,\dots,s_d\in\Delta$. Then the d-volume of $S=\operatorname{conv}\{s_0,\dots,s_d\}$ satisfies*

$$\operatorname{vol}(S) = |\det(s_0,\dots,s_d)|\operatorname{vol}(\Delta).$$

Proof. The $(d+1)$-volume of the parallelepiped spanned by $s_0,\dots,s_d\in\mathbb{R}^{d+1}$ is $|\det(s_0,\dots,s_d)|$. The volume of an n-simplex with vertices $v_0,\dots,v_n$ in $\mathbb{R}^n$ is $\frac{1}{n!}|\det(v_1-v_0,\dots,v_n-v_0)|$. Hence the volume of the $(d+1)$-simplex P with vertices $(0,s_0,\dots,s_d)$ is $\operatorname{vol}(P) = \frac{1}{(d+1)!}|\det(s_0,\dots,s_d)|$. Note that P is a pyramid over S of height $h=\frac{1}{\sqrt{d+1}}$. Thus $\operatorname{vol}(P)=\frac{h}{d+1}\operatorname{vol}(S)$. The volume of the regular d-simplex is $\operatorname{vol}(\Delta)=\frac{\sqrt{d+1}}{d!}$. The statement follows by combining these formulas. □

Example 16. Generalizing Examples 3 and 11, let G be the edge graph of an n-cube, such that $V=\{0,1\}^n$ and two points are adjacent if their Hamming distance is one.

a) If $\mathcal{C}\subseteq V$ has cardinality $|\mathcal{C}|=k$ and minimum distance 3, then $\mathbf{S}$ has 2^n vertices and volume $\operatorname{vol}(\mathbf{S})=2^{-kn}\operatorname{vol}(\Delta)$, whereas $\mathbf{M}$ has $k(2^n-1)+2^n-kn$ vertices and volume $\operatorname{vol}(\mathbf{M})=\frac{|\Sigma|}{2^n!}\operatorname{vol}(\Delta)\geq k!2^{-kn}\operatorname{vol}(\Delta)$.

b) If $\mathcal{C}$ is the set of all even-parity strings, then $\mathbf{S}$ has 2^n vertices and volume $\operatorname{vol}(\mathbf{S})=(n+1)^{-2^{n-1}}\operatorname{vol}(\Delta)$, whereas $\mathbf{M}$ has $2^{2^{n-1}}-1+2^{n-1}$ vertices and volume $\operatorname{vol}(\mathbf{M})=\frac{|\Sigma|}{2^n!}\operatorname{vol}(\Delta)\geq\binom{2^n}{2^{n-1}}^{-1}\operatorname{vol}(\Delta)$. For $n=2$ and $n=3$ we have $|\Sigma|=4$ and $|\Sigma|=720$. The next open case is $n=4$.

References

[1] G. Brightwell and P. Winkler. Counting linear extensions. *Order*, 8(3):225–242, 1991.

[2] G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics*, 29:321–347, 2015.

[3] R. Stanley. Two poset polytopes. *Discrete Comput. Geom.*, 1:9–23, 1986.

Representing Independence Models with Elementary Triplets

Jose M. Peña
ADIT, IDA, Linköping University, Sweden
jose.m.pena@liu.se

Abstract

An elementary triplet in an independence model represents a conditional independence statement between two singletons. It is known that these triplets can be used to represent the independence model unambiguously under some conditions. In this paper, we show how this representation helps performing efficiently some operations with independence models, such as finding the dominant triplets or a minimal independence map of an independence model, or computing the intersection or union of a pair of independence models.

1 Representation

Let V denote a finite set of elements. Subsets of V are denoted by upper-case letters, whereas the elements of V are denoted by lower-case letters. Given three sets $I, J, K \subseteq V$, the triplet $I \perp J|K$ denotes that I and J are conditionally independent given K. Given a set of triplets G, also known as an independence model, $I \perp_G J|K$ denotes that $I \perp J|K$ is in G. A triplet $I \perp J|K$ is called elementary if $|I| = |J| = 1$. We shall not distinguish between elements of V and singletons. We use IJ to denote $I \cup J$. Union has higher priority than set difference in expressions. Consider the following properties:

(CI0) $I \perp J|K \Leftrightarrow J \perp I|K$.
(CI1) $I \perp J|KL, I \perp K|L \Leftrightarrow I \perp JK|L$.
(CI2) $I \perp J|KL, I \perp K|JL \Rightarrow I \perp J|L, I \perp K|L$.
(CI3) $I \perp J|KL, I \perp K|JL \Leftarrow I \perp J|L, I \perp K|L$.

A set of triplets with the properties CI0-1/CI0-2/CI0-3 is also called a semigraphoid/graphoid/ compositional graphoid.[1] The CI0 property is also called symmetry property. The $\Rightarrow$ part of the CI1 property is also called contraction property, and

[1]For instance, the independencies in a probability distribution form a semigraphoid, while the independencies in a strictly positive probability distribution form a graphoid, and the independencies in a regular Gaussian distribution form a compositional graphoid.

the $\Leftarrow$ part corresponds to the so-called weak union and decomposition properties. The CI2 and CI3 properties are also called intersection and composition properties.[2] In addition, consider the following properties:

(ci0) $i \perp j|k \Leftrightarrow j \perp i|k$.
(ci1) $i \perp j|kL, i \perp k|L \Leftrightarrow i \perp k|jL, i \perp j|L$.
(ci2) $i \perp j|kL, i \perp k|jL \Rightarrow i \perp j|L, i \perp k|L$.
(ci3) $i \perp j|kL, i \perp k|jL \Leftarrow i \perp j|L, i \perp k|L$.

Note that CI2 and CI3 only differ in the direction of the implication. The same holds for ci2 and ci3.

Given a set of triplets $G = \{I \perp J|K\}$, let $\mathbb{P} = p(G) = \{i \perp j|M : I \perp_G J|K$ with $i \in I$, $j \in J$ and $K \subseteq M \subseteq (I \setminus i)(J \setminus j)K\}$. Given a set of elementary triplets $P = \{i \perp j|K\}$, let $\mathbb{G} = g(P) = \{I \perp J|K : i \perp_P j|M$ for all $i \in I$, $j \in J$ and $K \subseteq M \subseteq (I \setminus i)(J \setminus j)K\}$. The following two lemmas prove that there is a bijection between certain sets of triplets and certain sets of elementary triplets. The lemmas have been proven when G and P satisfy CI0-1 and ci0-1 [6, Proposition 1]. We extend them to the cases where G and P satisfy CI0-2/CI0-3 and ci0-2/ci0-3.

Lemma 1. *If G satisfies CI0-1/CI0-2/CI0-3 then (a) $\mathbb{P}$ satisfies ci0-1/ci0-2/ci0-3, (b) $G = g(\mathbb{P})$, and (c) $\mathbb{P} = \{i \perp j|K : i \perp_G j|K\}$.*

Proof. The proof of (c) is trivial. We now prove (a). That G satisfies C0 implies that $\mathbb{P}$ satisfies ci0 by definition of $\mathbb{P}$.

Proof of CI1 $\Rightarrow$ ci1

Since ci1 is symmetric, it suffices to prove the $\Rightarrow$ implication of ci1.

1. Assume that $i \perp_{\mathbb{P}} j|kL$.
2. Assume that $i \perp_{\mathbb{P}} k|L$.
3. Then, it follows from (1) and the definition of $\mathbb{P}$ that $i \perp_G j|kL$ or $I \perp_G J|M$ with $i \in I$, $j \in J$ and $M \subseteq kL \subseteq (I \setminus i)(J \setminus j)M$. Note that the latter case implies that $i \perp_G j|kL$ by CI1.
4. Then, $i \perp_G k|L$ by the same reasoning as in (3).
5. Then, $i \perp_G jk|L$ by CI1 on (3) and (4), which implies $i \perp_G k|jL$ and $i \perp_G j|L$ by CI1. Then, $i \perp_{\mathbb{P}} k|jL$ and $i \perp_{\mathbb{P}} j|L$ by definition of $\mathbb{P}$.

Proof of CI1-2 $\Rightarrow$ ci1-2

Assume that $i \perp_{\mathbb{P}} j|kL$ and $i \perp_{\mathbb{P}} k|jL$. Then, $i \perp_G j|kL$ and $i \perp_G k|jL$ by the same reasoning as in (3), which imply $i \perp_G j|L$ and $i \perp_G k|L$ by CI2. Then, $i \perp_{\mathbb{P}} j|L$ and $i \perp_{\mathbb{P}} k|L$ by definition of $\mathbb{P}$.

[2] Intersection is typically defined as $I \perp J|KL, I \perp K|JL \Rightarrow I \perp JK|L$. Note however that this and our definition are equivalent if CI1 holds. First, $I \perp JK|L$ implies $I \perp J|L$ and $I \perp K|L$ by CI1. Second, $I \perp J|L$ together with $I \perp K|JL$ imply $I \perp JK|L$ by CI1. Likewise, composition is typically defined as $I \perp JK|L \Leftarrow I \perp J|L, I \perp K|L$. Again, this and our definition are equivalent if CI1 holds. First, $I \perp JK|L$ implies $I \perp J|KL$ and $I \perp K|JL$ by CI1. Second, $I \perp K|JL$ together with $I \perp J|L$ imply $I \perp JK|L$ by CI1. In this paper, we will study sets of triplets that satisfy CI0-1, CI0-2 or CI0-3. So, the standard and our definitions are equivalent.

Proof of CI1-3 ⇒ ci1-3

Assume that $i \perp_{\mathbb{P}} j|L$ and $i \perp_{\mathbb{P}} k|L$. Then, $i \perp_G j|L$ and $i \perp_G k|L$ by the same reasoning as in (3), which imply $i \perp_G j|kL$ and $i \perp_G k|jL$ by CI3. Then, $i \perp_{\mathbb{P}} j|kL$ and $i \perp_{\mathbb{P}} k|jL$ by definition of $\mathbb{P}$.

Finally, we prove (b). Clearly, $G \subseteq g(\mathbb{P})$ by definition of $\mathbb{P}$. To see that $g(\mathbb{P}) \subseteq G$, note that $I \perp_{g(\mathbb{P})} J|K \Rightarrow I \perp_G J|K$ holds when $|I| = |J| = 1$. Assume as induction hypothesis that the result also holds when $2 < |IJ| < s$. Assume without loss of generality that $1 < |J|$. Let $J = J_1 J_2$ st $J_1, J_2 \neq \emptyset$ and $J_1 \cap J_2 = \emptyset$. Then, $I \perp_{g(\mathbb{P})} J_1|K$ and $I \perp_{g(\mathbb{P})} J_2|J_1K$ by ci1 and, thus, $I \perp_G J_1|K$ and $I \perp_G J_2|J_1K$ by the induction hypothesis, which imply $I \perp_G J|K$ by CI1. □

Lemma 2. *If P satisfies ci0-1/ci0-2/ci0-3 then (a) $\mathbb{G}$ satisfies CI0-1/CI0-2/CI0-3, (b) $P = p(\mathbb{G})$, and (c) $P = \{i \perp j|K : i \perp_{\mathbb{G}} j|K\}$.*

Proof. The proofs of (b) and (c) are trivial. We prove (a) below. That $\mathbb{P}$ satisfies ci0 implies that G satisfies C0 by definition of G.

Proof of ci1 ⇒ CI1

The ⇐ implication of CI1 is trivial. We prove below the ⇒ implication.

1. Assume that $I \perp_{\mathbb{G}} j|KL$.
2. Assume that $I \perp_{\mathbb{G}} K|L$.
3. Let $i \in I$. Note that if $i \not\perp_P j|M$ with $L \subseteq M \subseteq (I \setminus i)KL$ then (c) $i \not\perp_P j|kM$ with $k \in K \setminus M$, and (d) $i \not\perp_P j|KM$. To see (c), assume to the contrary that $i \perp_P j|kM$. This together with $i \perp_P k|M$ (which follows from (2) by definition of $\mathbb{G}$) imply that $i \perp_P j|M$ by ci1, which contradicts the assumption of $i \not\perp_P j|M$. To see (d), note that $i \not\perp_P j|M$ implies $i \not\perp_P j|kM$ with $k \in K \setminus M$ by (c), which implies $i \not\perp_P j|kk'M$ with $k' \in K \setminus kM$ by (c) again, and so on until the desired result is obtained.
4. Then, $i \perp_P j|M$ for all $i \in I$ and $L \subseteq M \subseteq (I \setminus i)KL$. To see it, note that $i \perp_P j|KM$ follows from (1) by definition of $\mathbb{G}$, which implies the desired result by (d) in (3).
5. $i \perp_P k|M$ for all $i \in I$, $k \in K$ and $L \subseteq M \subseteq (I \setminus i)(K \setminus k)L$ follows from (2) by definition of $\mathbb{G}$.
6. $i \perp_P k|jM$ for all $i \in I$, $k \in K$ and $L \subseteq M \subseteq (I \setminus i)(K \setminus k)L$ follows from ci1 on (4) and (5).
7. $I \perp_{\mathbb{G}} jK|L$ follows from (4)-(6) by definition of $\mathbb{G}$.

Therefore, we have proven above the ⇒ implication of CI1 when $|J| = 1$. Assume as induction hypothesis that the result also holds when $1 < |J| < s$. Let $J = J_1 J_2$ st $J_1, J_2 \neq \emptyset$ and $J_1 \cap J_2 = \emptyset$.

8. $I \perp_{\mathbb{G}} J_1|KL$ follows from $I \perp_{\mathbb{G}} J|KL$ by definition of $\mathbb{G}$.
9. $I \perp_{\mathbb{G}} J_2|J_1KL$ follows from $I \perp_{\mathbb{G}} J|KL$ by definition of $\mathbb{G}$.
10. $I \perp_{\mathbb{G}} J_1K|L$ by the induction hypothesis on (8) and $I \perp_{\mathbb{G}} K|L$.
11. $I \perp_{\mathbb{G}} JK|L$ by the induction hypothesis on (9) and (10).

Proof of ci1-2 ⇒ CI1-2

12. Assume that $I \perp_{\mathbb{G}} j|kL$ and $I \perp_{\mathbb{G}} k|jL$.
13. $i \perp_P j|kM$ and $i \perp_P k|jM$ for all $i \in I$ and $L \subseteq M \subseteq (I \setminus i)L$ follows from (12) by definition of $\mathbb{G}$.
14. $i \perp_P j|M$ and $i \perp_P k|M$ for all $i \in I$ and $L \subseteq M \subseteq (I \setminus i)L$ by ci2 on (13).
15. $I \perp_{\mathbb{G}} j|L$ and $I \perp_{\mathbb{G}} k|L$ follows from (14) by definition of $\mathbb{G}$.

Therefore, we have proven the result when $|J| = |K| = 1$. Assume as induction hypothesis that the result also holds when $2 < |JK| < s$. Assume without loss of generality that $1 < |J|$. Let $J = J_1J_2$ st $J_1, J_2 \neq \emptyset$ and $J_1 \cap J_2 = \emptyset$.

16. $I \perp_{\mathbb{G}} J_1|J_2KL$ and $I \perp_{\mathbb{G}} J_2|J_1KL$ by CI1 on $I \perp_{\mathbb{G}} J|KL$.
17. $I \perp_{\mathbb{G}} J_1|J_2L$ and $I \perp_{\mathbb{G}} J_2|J_1L$ by the induction hypothesis on (16) and $I \perp_{\mathbb{G}} K|JL$.
18. $I \perp_{\mathbb{G}} J|L$ by the induction hypothesis on (17).
19. $I \perp_{\mathbb{G}} K|L$ by CI1 on (18) and $I \perp_{\mathbb{G}} K|JL$.

Proof of ci1-3 ⇒ CI1-3

20. Assume that $I \perp_{\mathbb{G}} j|L$ and $I \perp_{\mathbb{G}} k|L$.
21. $i \perp_P j|M$ and $i \perp_P k|M$ for all $i \in I$ and $L \subseteq M \subseteq (I \setminus i)L$ follows from (20) by definition of $\mathbb{G}$.
22. $i \perp_P j|kM$ and $i \perp_P k|jM$ for all $i \in I$ and $L \subseteq M \subseteq (I \setminus i)L$ by ci3 on (21).
23. $I \perp_{\mathbb{G}} j|kL$ and $I \perp_{\mathbb{G}} k|jL$ follows from (22) by definition of $\mathbb{G}$.

Therefore, we have proven the result when $|J| = |K| = 1$. Assume as induction hypothesis that the result also holds when $2 < |JK| < s$. Assume without loss of generality that $1 < |J|$. Let $J = J_1J_2$ st $J_1, J_2 \neq \emptyset$ and $J_1 \cap J_2 = \emptyset$.

24. $I \perp_{\mathbb{G}} J_1|L$ by CI1 on $I \perp_{\mathbb{G}} J|L$.
25. $I \perp_{\mathbb{G}} J_2|J_1L$ by CI1 on $I \perp_{\mathbb{G}} J|L$.
26. $I \perp_{\mathbb{G}} K|J_1L$ by the induction hypothesis on (24) and $I \perp_{\mathbb{G}} K|L$.
27. $I \perp_{\mathbb{G}} K|JL$ by the induction hypothesis on (25) and (26).
28. $I \perp_{\mathbb{G}} JK|L$ by CI1 on (27) and $I \perp_{\mathbb{G}} J|L$.
29. $I \perp_{\mathbb{G}} J|KL$ and $I \perp_{\mathbb{G}} K|JL$ by CI1 on (28).

□

The following two lemmas generalize Lemmas 1 and 2 by removing the assumptions about G and P.

Lemma 3. *Let G^* denote the CI0-1/CI0-2/CI0-3 closure of G, and let $\mathbb{P}^*$ denote the ci0-1/ci0-2/ci0-3 closure of $\mathbb{P}$. Then, $\mathbb{P}^* = p(G^*)$, $G^* = g(\mathbb{P}^*)$ and $\mathbb{P}^* = \{i \perp j|K : i \perp_{G^*} j|K\}$.*

Proof. Clearly, $G \subseteq g(\mathbb{P}^*)$ and, thus, $G^* \subseteq g(\mathbb{P}^*)$ because $g(\mathbb{P}^*)$ satisfies CI0-1/CI0-2/CI0-3 by Lemma 2. Clearly, $\mathbb{P} \subseteq p(G^*)$ and, thus, $\mathbb{P}^* \subseteq p(G^*)$ because $p(G^*)$ satisfies ci0-1/ci0-2/ci0-3 by Lemma 1. Then, $G^* \subseteq g(\mathbb{P}^*) \subseteq g(p(G^*))$ and $\mathbb{P}^* \subseteq p(G^*) \subseteq p(g(\mathbb{P}^*))$. Then, $G^* = g(\mathbb{P}^*)$ and $\mathbb{P}^* = p(G^*)$, because $G^* = g(p(G^*))$ and $\mathbb{P}^* = p(g(\mathbb{P}^*))$ by Lemmas 1 and 2. Finally, that $\mathbb{P}^* = \{i \perp j|K : i \perp_{G^*} j|K\}$ is now trivial. □

Lemma 4. *Let P^* denote the ci0-1/ci0-2/ci0-3 closure of P, and let $\mathbb{G}^*$ denote the CI0-1/CI0-2/CI0-3 closure of $\mathbb{G}$. Then, $\mathbb{G}^* = g(P^*)$, $P^* = p(\mathbb{G}^*)$ and $P^* = \{i \perp j|K : i \perp_{\mathbb{G}^*} j|K\}$.*

Proof. Clearly, $P \subseteq p(\mathbb{G}^*)$ and, thus, $P^* \subseteq p(\mathbb{G}^*)$ because $p(\mathbb{G}^*)$ satisfies ci0-1/ci0-2/ci0-3 by Lemma 1. Clearly, $\mathbb{G} \subseteq g(P^*)$ and, thus, $\mathbb{G}^* \subseteq g(P^*)$ because $g(P^*)$ satisfies CI0-1/CI0-2/CI0-3 by Lemma 2. Then, $P^* \subseteq p(\mathbb{G}^*) \subseteq p(g(P^*))$ and $\mathbb{G}^* \subseteq g(P^*) \subseteq g(p(\mathbb{G}^*))$. Then, $P^* = p(\mathbb{G}^*)$ and $\mathbb{G}^* = g(P^*)$, because $P^* = p(g(P^*))$ and $\mathbb{G}^* = g(p(\mathbb{G}^*))$ by Lemmas 1 and 2. Finally, that $P^* = \{i \perp j|K : i \perp_{\mathbb{G}^*} j|K\}$ is now trivial. □

The parts (a) of Lemmas 1 and 2 imply that every set of triplets G satisfying CI0-1/CI0-2/CI0-3 can be paired to a set of elementary triplets P satisfying ci0-1/ci0-2/ci0-3, and vice versa. The pairing is actually a bijection, due to the parts (b) of the lemmas. Thanks to this bijection, we can use $\mathbb{P}$ to represent G. This is in general a much more economical representation: If $|V| = n$, then there up to 4^n triplets,[3] whereas there are $n^2 \cdot 2^{n-2}$ elementary triplets at most. We can reduce further the size of the representation by iteratively removing from $\mathbb{P}$ an elementary triplet that follows from two others by ci0-1/ci0-2/ci0-3. Note that $\mathbb{P}$ is an unique representation of G but the result of the removal process is not in general, as ties may occur during the process.

Likewise, Lemmas 3 and 4 imply that there is a bijection between the CI0-1/CI0-2/CI0-3 closures of sets of triplets and the ci0-1/ci0-2/ci0-3 closures of sets of elementary triplets. Thanks to this bijection, we can use $\mathbb{P}^*$ to represent G^*. Note that $\mathbb{P}^*$ is obtained by ci0-1/ci0-2/ci0-3 closing $\mathbb{P}$, which is obtained from G. So, there is no need to CI0-1/CI0-2/CI0-3 close G and so produce G^*. Whether closing $\mathbb{P}$ can be done faster than closing G on average is an open question. In the worst-case scenario, both imply applying the corresponding properties a number of times exponential in $|V|$ [7]. We can avoid this problem by simply using $\mathbb{P}$ to represent G^*, because $\mathbb{P}$ is the result of running the removal process outline above on $\mathbb{P}^*$. All the results in the sequel assume that G and P satisfy CI0-1/CI0-2/CI0-3 and ci0-1/ci0-2/ci0-3. Thanks to Lemmas 3 and 4, these assumptions can be dropped by replacing G, P, $\mathbb{G}$ and $\mathbb{P}$ in the results below with G^*, P^*, $\mathbb{G}^*$ and $\mathbb{P}^*$.

Let $I = i_1 \dots i_m$ and $J = j_1 \dots j_n$. In order to decide whether $I \perp_{\mathbb{G}} J|K$, the definition of $\mathbb{G}$ implies checking whether $m \cdot n \cdot 2^{(m+n-2)}$ elementary triplets are in P. The following lemma simplifies this for when P satisfies ci0-1, as it implies checking $m \cdot n$ elementary triplets. For when P satisfies ci0-2 or ci0-3, the lemma simplifies the

[3] A triplet can be represented as a n-tuple whose entries state if the corresponding node is in the first, second, third or none set of the triplet.

decision even further as the conditioning sets of the elementary triplets checked have all the same size or form.

Lemma 5. *Let* $\mathbb{H}_1 = \{I \perp J|K : i_s \perp_P j_t|i_1 \dots i_{s-1}j_1 \dots j_{t-1}K$ *for all* $1 \leq s \leq m$ *and* $1 \leq t \leq n\}$, $\mathbb{H}_2 = \{I \perp J|K : i \perp_P j|(I \setminus i)(J \setminus j)K$ *for all* $i \in I$ *and* $j \in J\}$, *and* $\mathbb{H}_3 = \{I \perp J|K : i \perp_P j|K$ *for all* $i \in I$ *and* $j \in J\}$. *If* P *satisfies ci0-1, then* $\mathbb{G} = \mathbb{H}_1$. *If* P *satisfies ci0-2, then* $\mathbb{G} = \mathbb{H}_2$. *If* P *satisfies ci0-3, then* $\mathbb{G} = \mathbb{H}_3$.

Proof. Proof for ci0-1

It suffices to prove that $\mathbb{H}_1 \subseteq \mathbb{G}$, because it is clear that $\mathbb{G} \subseteq \mathbb{H}_1$. Assume that $I \perp_{\mathbb{H}_1} J|K$. Then, $i_s \perp_P j_t|i_1 \dots i_{s-1}j_1 \dots j_{t-1}K$ and $i_s \perp_P j_{t+1}|i_1 \dots i_{s-1}j_1 \dots j_tK$ by definition of $\mathbb{H}_1$. Then, $i_s \perp_P j_{t+1}|i_1 \dots i_{s-1}j_1 \dots j_{t-1}K$ and $i_s \perp_P j_t|i_1 \dots i_{s-1}$ $j_1 \dots j_{t-1}j_{t+1}K$ by ci1. Then, $i_s \perp_{\mathbb{G}} j_{t+1}|i_1 \dots i_{s-1}j_1 \dots j_{t-1}K$ and $i_s \perp_{\mathbb{G}} j_t|i_1 \dots i_{s-1}$ $j_1 \dots j_{t-1}j_{t+1}K$ by definition of $\mathbb{G}$. By repeating this reasoning, we can then conclude that $i_s \perp_{\mathbb{G}} j_{\sigma(t)}|i_1 \dots i_{s-1}j_{\sigma(1)} \dots j_{\sigma(t-1)}K$ for any permutation σ of the set $\{1 \dots n\}$. By following an analogous reasoning for i_s instead of j_t, we can then conclude that $i_{\varsigma(s)} \perp_{\mathbb{G}} j_{\sigma(t)}|i_{\varsigma(1)} \dots i_{\varsigma(s-1)}j_{\sigma(1)} \dots j_{\sigma(t-1)}K$ for any permutations σ and ς of the sets $\{1 \dots n\}$ and $\{1 \dots m\}$. This implies the desired result by definition of $\mathbb{G}$.

Proof for ci0-2

It suffices to prove that $\mathbb{H}_2 \subseteq \mathbb{G}$, because it is clear that $\mathbb{G} \subseteq \mathbb{H}_2$. Note that $\mathbb{G}$ satisfies CI0-2 by Lemma 2. Assume that $I \perp_{\mathbb{H}_2} J|K$.

1. $i_1 \perp_{\mathbb{G}} j_1|(I \setminus i_1)(J \setminus j_1)K$ and $i_1 \perp_{\mathbb{G}} j_2|(I \setminus i_1)(J \setminus j_2)K$ follow from $i_1 \perp_P j_1|(I \setminus i_1)(J \setminus j_1)K$ and $i_1 \perp_P j_2|(I \setminus i_1)(J \setminus j_2)K$ by definition of $\mathbb{G}$.
2. $i_1 \perp_{\mathbb{G}} j_1|(I \setminus i_1)(J \setminus j_1j_2)K$ by CI2 on (1), which together with (1) imply $i_1 \perp_{\mathbb{G}} j_1j_2|(I \setminus i_1)(J \setminus j_1j_2)K$ by CI1.
3. $i_1 \perp_{\mathbb{G}} j_3|(I \setminus i_1)(J \setminus j_3)K$ follows from $i_1 \perp_P j_3|(I \setminus i_1)(J \setminus j_3)K$ by definition of $\mathbb{G}$.
4. $i_1 \perp_{\mathbb{G}} j_1j_2|(I \setminus i_1)(J \setminus j_1j_2j_3)K$ by CI2 on (2) and (3), which together with (3) imply $i_1 \perp_{\mathbb{G}} j_1j_2j_3|(I \setminus i_1)(J \setminus j_1j_2j_3)K$ by CI1.

By continuing with the reasoning above, we can conclude that $i_1 \perp_{\mathbb{G}} J|(I \setminus i_1)K$. By an analogous reasoning, we can conclude that $i_1i_2 \perp_{\mathbb{G}} J|(I \setminus i_1i_2)K$, $i_1i_2i_3 \perp_{\mathbb{G}} J|(I \setminus i_1i_2i_3)K$ and so on until the desired is obtained.

Proof for ci0-3

It suffices to prove that $\mathbb{H}_3 \subseteq \mathbb{G}$, because it is clear that $\mathbb{G} \subseteq \mathbb{H}_3$. Note that $\mathbb{G}$ satisfies CI0-3 by Lemma 2. Assume that $I \perp_{\mathbb{H}_3} J|K$.

1. $i_1 \perp_{\mathbb{G}} j_1|K$ and $i_1 \perp_{\mathbb{G}} j_2|K$ follow from $i_1 \perp_P j_1|K$ and $i_1 \perp_P j_2|K$ by definition of $\mathbb{G}$.
2. $i_1 \perp_{\mathbb{G}} j_1|j_2K$ by CI3 on (1), which together with (1) imply $i_1 \perp_{\mathbb{G}} j_1j_2|K$ by CI1.
3. $i_1 \perp_{\mathbb{G}} j_3|K$ follows from $i_1 \perp_P j_3|K$ by definition of $\mathbb{G}$.
4. $i_1 \perp_{\mathbb{G}} j_1j_2|j_3K$ by CI3 on (2) and (3), which together with (3) imply $i_1 \perp_{\mathbb{G}} j_1j_2j_3|K$ by CI1.

By continuing with the reasoning above, we can conclude that $i_1 \perp_{\mathbb{G}} J|K$. By an analogous reasoning, we can conclude that $i_1 i_2 \perp_{\mathbb{G}} J|K$, $i_1 i_2 i_3 \perp_{\mathbb{G}} J|K$ and so on until the desired result is obtained. □

We are not the first to use some distinguished triplets of G to represent it. However, most other works use dominant triplets for this purpose [1, 4, 5, 9]. The following lemma shows how to find these triplets with the help of $\mathbb{P}$. A triplet $I \perp J|K$ dominates another triplet $I' \perp J'|K'$ if $I' \subseteq I$, $J' \subseteq J$ and $K \subseteq K' \subseteq (I \setminus I')(J \setminus J')K$. Given a set of triplets, a triplet in the set is called dominant if no other triplet in the set dominates it.

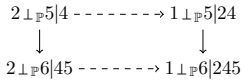
Lemma 6. *If G satisfies CI0-1, then $I \perp J|K$ is a dominant triplet in G iff $I = i_1 \dots i_m$ and $J = j_1 \dots j_n$ are two maximal sets st $i_s \perp_{\mathbb{P}} j_t | i_1 \dots i_{s-1} j_1 \dots j_{t-1} K$ for all $1 \leq s \leq m$ and $1 \leq t \leq n$ and, for all $k \in K$, $i_s \not\perp_{\mathbb{P}} k | i_1 \dots i_{s-1} J(K \setminus k)$ and $k \not\perp_{\mathbb{P}} j_t | I j_1 \dots j_{t-1}(K \setminus k)$ for some $1 \leq s \leq m$ and $1 \leq t \leq n$. If G satisfies CI0-2, then $I \perp J|K$ is a dominant triplet in G iff I and J are two maximal sets st $i \perp_{\mathbb{P}} j|(I \setminus i)(J \setminus j)K$ for all $i \in I$ and $j \in J$ and, for all $k \in K$, $i \not\perp_{\mathbb{P}} k|(I \setminus i)J(K \setminus k)$ and $k \not\perp_{\mathbb{P}} j|I(J \setminus j)(K \setminus k)$ for some $i \in I$ and $j \in J$. If G satisfies CI0-3, then $I \perp J|K$ is a dominant triplet in G iff I and J are two maximal sets st $i \perp_{\mathbb{P}} j|K$ for all $i \in I$ and $j \in J$ and, for all $k \in K$, $i \not\perp_{\mathbb{P}} k|K \setminus k$ and $k \not\perp_{\mathbb{P}} j|K \setminus k$ for some $i \in I$ and $j \in J$.*

Proof. We proof the lemma for when G satisfies CI0-1. The other two cases can be proven in much the same way. To see the if part, note that $I \perp_G J|K$ by Lemmas 1 and 5. Moreover, assume to the contrary that there is a triplet $I' \perp_G J'|K'$ that dominates $I \perp_G J|K$. Consider the following two cases: $K' = K$ and $K' \subset K$. In the first case, CI1 on $I' \perp_G J'|K'$ implies that $I i_{m+1} \perp_G J|K$ or $I \perp_G J j_{n+1}|K$ with $i_{m+1} \in I' \setminus I$ and $j_{n+1} \in J' \setminus J$. Assume the latter without loss of generality. Then, CI1 implies that $i_s \perp_{\mathbb{P}} j_t | i_1 \dots i_{s-1} j_1 \dots j_{t-1} K$ for all $1 \leq s \leq m$ and $1 \leq t \leq n+1$. This contradicts the maximality of J. In the second case, CI1 on $I' \perp_G J'|K'$ implies that $Ik \perp_G J|K \setminus k$ or $I \perp_G Jk|K \setminus k$ with $k \in K$. Assume the latter without loss of generality. Then, CI1 implies that $i_s \perp_{\mathbb{P}} k | i_1 \dots i_{s-1} J(K \setminus k)$ for all $1 \leq s \leq m$, which contradicts the assumptions of the lemma.

To see the only if part, note that CI1 implies that $i_s \perp_{\mathbb{P}} j_t | i_1 \dots i_{s-1} j_1 \dots j_{t-1} K$ for all $1 \leq s \leq m$ and $1 \leq t \leq n$. Moreover, assume to the contrary that for some $k \in K$, $i_s \perp_{\mathbb{P}} k | i_1 \dots i_{s-1} J(K \setminus k)$ for all $1 \leq s \leq m$ or $k \perp_{\mathbb{P}} j_t | I j_1 \dots j_{t-1}(K \setminus k)$ for all $1 \leq t \leq n$. Assume the latter without loss of generality. Then, $Ik \perp_G J|K \setminus k$ by Lemmas 1 and 5, which implies that $I \perp_G J|K$ is not a dominant triplet in G, which is a contradiction. Finally, note that I and J must be maximal sets satisfying the properties proven in this paragraph because, otherwise, the previous paragraph implies that there is a triplet in G that dominates $I \perp_G J|K$. □

Inspired by [7], if G satisfies CI0-1 then we represent $\mathbb{P}$ as a DAG. The nodes of the DAG are the elementary triplets in $\mathbb{P}$ and the edges of the DAG are $\{i \perp_{\mathbb{P}} k|L \rightarrow i \perp_{\mathbb{P}} j|kL\} \cup \{k \perp_{\mathbb{P}} j|L \dashrightarrow i \perp_{\mathbb{P}} j|kL\}$. See Figure 1 for an example. For the sake of readability, the DAG in the figure does not include symmetric elementary triplets. That is, the complete DAG can be obtained by adding a second copy of the DAG

in the figure, replacing every node $i \perp_{\mathbb{P}} j|K$ in the copy with $j \perp_{\mathbb{P}} i|K$, and replacing every edge $\rightarrow$ in the copy with $\dashrightarrow$. We say that a subgraph over $m \cdot n$ nodes of the DAG is a grid if there is a bijection between the nodes of the subgraph and the labels $\{v_{s,t} : 1 \leq s \leq m, 1 \leq t \leq n\}$ st the edges of the subgraph are $\{v_{s,t} \rightarrow v_{s,t+1} : 1 \leq s \leq m, 1 \leq t < n\} \cup \{v_{s,t} \dashrightarrow v_{s+1,t} : 1 \leq s < m, 1 \leq t \leq n\}$. For instance, the following subgraph of the DAG in Figure 1 is a grid.



The following lemma is an immediate consequence of Lemmas 1 and 5.

Lemma 7. *Let G satisfy CI0-1, and let $I = i_1 \dots i_m$ and $J = j_1 \dots j_n$. If the subgraph of the DAG representation of $\mathbb{P}$ induced by the set of nodes $\{i_s \perp_{\mathbb{P}} j_t | i_1 \dots i_{s-1} j_1 \dots j_{t-1} K : 1 \leq s \leq m, 1 \leq t \leq n\}$ is a grid, then $I \perp_G J|K$.*

Thanks to Lemmas 6 and 7, finding dominant triplets can now be reformulated as finding maximal grids in the DAG. Note that this is a purely graphical characterization. For instance, the DAG in Figure 1 has 18 maximal grids: The subgraphs induced by the set of nodes $\{\sigma(s) \perp_{\mathbb{P}} \varsigma(t) | \sigma(1) \dots \sigma(s-1) \varsigma(1) \dots \varsigma(t-1) : 1 \leq s \leq 2, 1 \leq t \leq 3\}$ where σ and ς are permutations of $\{1, 2\}$ and $\{4, 5, 6\}$, and the set of nodes $\{\pi(s) \perp_{\mathbb{P}} 4 | \pi(1) \dots \pi(s-1) : 1 \leq s \leq 3\}$ where π is a permutation of $\{1, 2, 3\}$. These grids correspond to the dominant triplets $12 \perp_G 456|\varnothing$ and $123 \perp_G 4|\varnothing$.

2 Operations

In this section, we discuss some operations with independence models that can efficiently be performed with the help of $\mathbb{P}$. See [2, 3] for how to perform these operations efficiently when independence models are represented by their dominant triplets.

2.1 Membership

We want to check whether $I \perp_G J|K$, where G denotes a set of triplets satisfying CI0-1/CI0-2/CI0-3. Recall that G can be obtained from $\mathbb{P}$ by Lemma 1. Recall also that $\mathbb{P}$ satisfies ci0-1/ci0-2/ci0-3 by Lemma 1 and, thus, Lemma 5 applies to $\mathbb{P}$, which simplifies producing G from $\mathbb{P}$. Specifically if G satisfies CI0-1, then we can check whether $I \perp_G J|K$ with $I = i_1 \dots i_m$ and $J = j_1 \dots j_n$ by checking whether $i_s \perp_{\mathbb{P}} j_t | i_1 \dots i_{s-1} j_1 \dots j_{t-1} K$ for all $1 \leq s \leq m$ and $1 \leq t \leq n$. Thanks to Lemma 7, this solution can also be reformulated as checking whether the DAG representation of $\mathbb{P}$ contains a suitable grid. Likewise, if G satisfies CI0-2, then we can check whether $I \perp_G J|K$ by checking whether $i \perp_{\mathbb{P}} j | (I \setminus i)(J \setminus j)K$ for all $i \in I$ and $j \in J$. Finally, if G satisfies CI0-3, then we can check whether $I \perp_G J|K$ by checking whether $i \perp_{\mathbb{P}} j|K$ for all $i \in I$ and $j \in J$. Note that in the last two cases, we only need to check elementary triplets with conditioning sets of a specific length or form.

2.2 Minimal Independence Map

We say that a DAG D is a minimal independence map (MIM) of a set of triplets G relative to an ordering σ of the elements in V if (i) $I \perp_D J|K \Rightarrow I \perp_G J|K$,[4] (ii) removing any edge from D makes it cease to satisfy condition (i), and (iii) the edges of D are of the form $\sigma(s) \rightarrow \sigma(t)$ with $s < t$. If G satisfies CI0-1, then D can be built by setting $Pa_D(\sigma(s))$[5] for all $1 \leq s \leq |V|$ to a minimal subset of $\sigma(1) \ldots \sigma(s-1)$ st $\sigma(s) \perp_G \sigma(1) \ldots \sigma(s-1) \setminus Pa_D(\sigma(s)) | Pa_D(\sigma(s))$ [8, Theorem 9]. Thanks to Lemma 7, building a MIM of G relative to σ can now be reformulated as finding, for all $1 \leq s \leq |V|$, a longest grid in the DAG representation of $\mathbb{P}$ that is of the form $\sigma(s) \perp_{\mathbb{P}} j_1 | \sigma(1) \ldots \sigma(s-1) \setminus j_1 \ldots j_n \rightarrow \sigma(s) \perp_{\mathbb{P}} j_2 | \sigma(1) \ldots \sigma(s-1) \setminus j_2 \ldots j_n \rightarrow \ldots \rightarrow \sigma(s) \perp_{\mathbb{P}} j_n | \sigma(1) \ldots \sigma(s-1) \setminus j_n$, or $j_1 \perp_{\mathbb{P}} \sigma(s) | \sigma(1) \ldots \sigma(s-1) \setminus j_1 \ldots j_n \dashrightarrow j_2 \perp_{\mathbb{P}} \sigma(s) | \sigma(1) \ldots \sigma(s-1) \setminus j_2 \ldots j_n \dashrightarrow \ldots \dashrightarrow j_n \perp_{\mathbb{P}} \sigma(s) | \sigma(1) \ldots \sigma(s-1) \setminus j_n$ with $j_1 \ldots j_n \subseteq \sigma(1) \ldots \sigma(s-1)$. Then, we set $Pa_D(\sigma(s))$ to $\sigma(1) \ldots \sigma(s-1) \setminus j_1 \ldots j_n$.

We say that a DAG D is a perfect map (PM) of a set of triplets G if $I \perp_D J|K \Leftrightarrow I \perp_G J|K$. We can check whether G has a PM with the help of $\mathbb{P}$ as follows: G has a PM iff $PM(\emptyset, \emptyset)$ returns true, where

$PM(Visited, Marked)$

if $Visited = V$ then
  if all the nodes in the DAG representation of $\mathbb{P}$ are in $Marked$ then
return true and stop
else
  for each node $i \in V \setminus Visited$ do
    for each longest grid in the DAG representation of $\mathbb{P}$ that is of the form
    $i \perp_{\mathbb{P}} j_1 | Visited \setminus j_1 \ldots j_n \rightarrow i \perp_{\mathbb{P}} j_2 | Visited \setminus j_2 \ldots j_n \rightarrow \ldots \rightarrow i \perp_{\mathbb{P}} j_n | Visited \setminus j_n$ or
    $j_1 \perp_{\mathbb{P}} i | Visited \setminus j_1 \ldots j_n \dashrightarrow j_2 \perp_{\mathbb{P}} i | Visited \setminus j_2 \ldots j_n \dashrightarrow \ldots \dashrightarrow j_n \perp_{\mathbb{P}} i | Visited \setminus j_n$ with
    $j_1 \ldots j_n \subseteq Visited$ do
      $PM(Visited \cup \{i\},$
      $Marked \cup p(\{i \perp_G j_1 \ldots j_n | Visited \setminus j_1 \ldots j_n\}) \cup p(\{j_1 \ldots j_n \perp_G i | Visited \setminus j_1 \ldots j_n\}))$

2.3 Inclusion

Let G and G' denote two sets of triplets satisfying CI0-1/CI0-2/CI0-3. We can check whether $G \subseteq G'$ by checking whether $\mathbb{P} \subseteq \mathbb{P}'$. If the DAG representations of $\mathbb{P}$ and $\mathbb{P}'$ are available, then we can answer the inclusion question by checking whether the former is a subgraph of the latter.

2.4 Intersection

Let G and G' denote two sets of triplets satisfying CI0-1/CI0-2/CI0-3. Note that $G \cap G'$ satisfies CI0-1/CI0-2/CI0-3. Likewise, $\mathbb{P} \cap \mathbb{P}'$ satisfies ci0-1/ci0-2/ci0-3. We

[4] $I \perp_D J|K$ stands for I and J are d-separated in D given K.
[5] $Pa_D(\sigma(s))$ denotes the parents of $\sigma(s)$ in D.

can represent $G \cap G$ by $\mathbb{P} \cap \mathbb{P}'$. To see it, note that $I \perp_{G \cap G'} J|K$ iff $i \perp_{\mathbb{P}} j|M$ and $i \perp_{\mathbb{P}'} j|M$ for all $i \in I$, $j \in J$, and $K \subseteq M \subseteq (I \setminus i)(J \setminus j)K$. If the DAG representations of $\mathbb{P}$ and $\mathbb{P}'$ are available, then we can represent $G \cap G$ by the subgraph of either of them induced by the nodes that are in both of them.

2.5 Union

Let G and G' denote two sets of triplets satisfying CI0-1/CI0-2/CI0-3. Note that $G \cup G'$ may not satisfy CI0-1/CI0-2/CI0-3. For instance, let $G = \{x \perp y|z, y \perp x|z\}$ and $G' = \{x \perp z|\emptyset, z \perp x|\emptyset\}$. We can solve this problem by simply adding an auxiliary element e (respectively e') to the conditioning set of every triplet in G (respectively G'). In the previous example, $G = \{x \perp y|ze, y \perp x|ze\}$ and $G' = \{x \perp z|e', z \perp x|e'\}$. Now, we can represent $G \cup G'$ by first adding the auxiliary element e (respectively e') to the conditioning set of every elementary triplet in $\mathbb{P}$ (respectively $\mathbb{P}'$) and, then, taking $\mathbb{P} \cup \mathbb{P}'$. This solution has advantages and disadvantages. The main advantage is that we represent $G \cup G'$ exactly. One of the disadvantages is that the same elementary triplet may appear twice in the representation, i.e. with e and e' in the conditioning set. Another disadvantage is that we need to modify slightly the procedures described above for building MIMs, and checking membership and inclusion. We believe that the advantage outweighs the disadvantages.

If the solution above is not satisfactory, then we have two options: Representing a minimal superset or a maximal superset of $G \cup G'$ satisfying CI0-1/CI0-2/CI0-3. Note that the minimal superset of $G \cup G'$ satisfying CI0-1/CI0-2/CI0-3 is unique because, otherwise, the intersection of any two such supersets is a superset of $G \cup G'$ that satisfies CI0-1/CI0-2/CI0-3, which contradicts the minimality of the original supersets. On the other hand, the maximal subset of $G \cup G'$ satisfying CI0-1/CI0-2/CI0-3 is not unique. For instance, let $G = \{x \perp y|z, y \perp x|z\}$ and $G' = \{x \perp z|\emptyset, z \perp x|\emptyset\}$. We can represent the minimal superset of $G \cup G'$ satisfying CI0-1/CI0-2/CI0-3 by the ci0-1/ci0-2/ci0-3 closure of $\mathbb{P} \cup \mathbb{P}'$. Clearly, this representation represents a superset of $G \cup G'$. Moreover, the superset satisfies CI0-1/CI0-2/CI0-3 by Lemma 2. Minimality follows from the fact that removing any elementary triplet from the representation implies not representing some triplet in $G \cup G'$ by Lemma 1. Note that the DAG representation of $G \cup G'$ is not the union of the DAG representations of $\mathbb{P}$ and $\mathbb{P}'$, because we first have to close $\mathbb{P} \cup \mathbb{P}'$ under ci0-1/ci0-2/ci0-3. We can represent a maximal subset of $G \cup G'$ satisfying CI0-1/CI0-2/CI0-3 by a maximal subset U of $\mathbb{P} \cup \mathbb{P}'$ that is closed under ci0-1/ci0-2/ci0-3 and st every triplet represented by U is in $G \cup G'$. Recall that we can efficiently check the latter as shown above. In fact, we do not need to check it for every triplet but only for the dominant triplets. Recall that these can efficiently be obtained from U as shown in the previous section.

3 Discussion

In this work, we have proposed to represent semigraphoids, graphoids and compositional graphoids by their elementary triplets. We have also shown how this representation helps performing efficiently some common operations between independence

models. Whether this implies a gain of efficiency compared to other representations (e.g. dominant triplets) is a question for future research.

References

[1] Marco Baioletti, Giuseppe Busanello, and Barbara Vantaggi. Conditional independence structure and its closure: Inferential rules and algorithms. *International Journal of Approximate Reasoning*, 50(7):1097–1114, 2009.

[2] Marco Baioletti, Giuseppe Busanello, and Barbara Vantaggi. Acyclic directed graphs representing independence models. *International Journal of Approximate Reasoning*, 52(1):2 – 18, 2011.

[3] Marco Baioletti, Davide Petturiti, and Barbara Vantaggi. Qualitative combination of independence models. In *Proceedings of the 12th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 37–48, 2013.

[4] Peter de Waal and Linda C. van der Gaag. Stable independence and complexity of representation. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 112–119, 2004.

[5] Stavros Lopatatzidis and Linda C. van der Gaag. Computing concise representations of semi-graphoid independency models. In *Proceedings of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 290–300, 2015.

[6] Frantisek Matús. Ascending and descending conditional independence relations. In *Proceedings of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 189–200, 1992.

[7] Frantisek Matús. Lengths of semigraphoid inferences. *Annals of Mathematics and Artificial Intelligence*, 35:287–294, 2002.

[8] Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.

[9] Milan Studený. Complexity of structural models. In *Proceedings of the Joint Session of the 6th Prague Conference on Asymptotic Statistics and the 13th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 521–528, 1998.

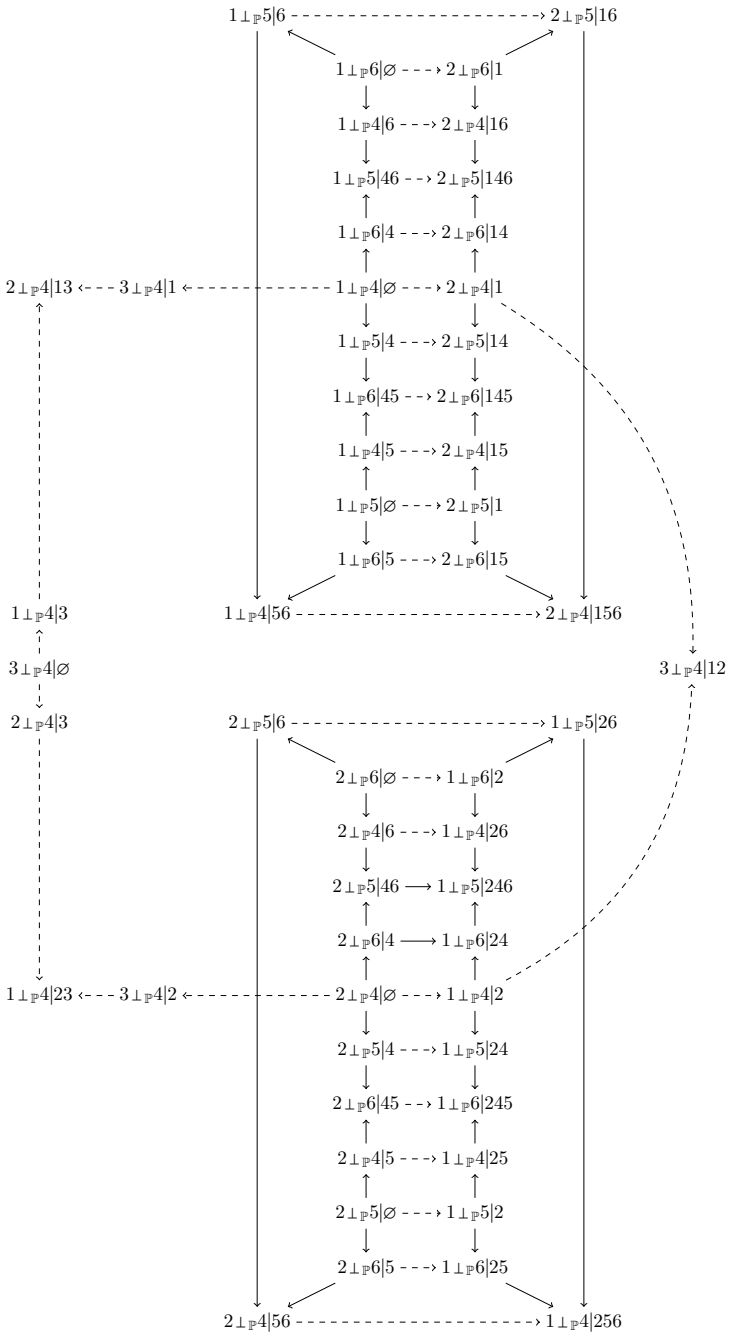


Figure 1: DAG representation of $\mathbb{P}$ (up to symmetry).

Decomposition of Markov Kernels

Paolo Perrone and Nihat Ay
Max Planck Institute for Mathematics in the Sciences
Leipzig, Germany
perrone@mis.mpg.de

Abstract

The decomposition of information into unique, shared, and synergetic parts is an open, active problem in the theory of complex systems. Most approaches to the problem are based on information theory, and propose decompositions of mutual information between inputs and outputs in several ways, none of which is generally accepted (yet).

We propose a new point of view on the topic. We model a multi-input channel as a Markov kernel. We can decompose the kernel into a series of single input nodes which represent single node information; pairwise interactions; and in general onto n-node interactions, which form a hierarchical structure.

We consider three different ways to do the decomposition:

1. Linearly (Section 1), using orthogonal projectors w.r.t. the L^2 inner product defined by an input probability distribution.
2. Algebraically (Section 2), using algebraic-statistical quantities which quantify the amount of interaction.
3. Geometrically (Section 3), minimizing the KL divergence between different exponential families, or equivalently, maximizing the entropy with constrains on the marginals.

Under particular conditions, the three approaches are similar. Advantages and disadvantages are outlined at the end.

Keywords: Synergy, Redundancy, Markov Kernels, Hilbert Spaces, Decomposition, Projections, Divergences, Interactions

Introduction

In complex systems like biological networks, for example neural networks, a basic principle is that their functioning is based on the correlation and interaction of their different parts. While correlation between two sources is well understood, and can be quantified by Shannon's mutual information (see for example [15]), there is still no generally accepted theory for interactions of three nodes or more. If we label one of the nodes as "output", the problem is equivalent to determine how much two (or more) input nodes interact to yield the output.

History

There are a number of important works which address the topic, but the problem is still considered open. The first generalization of mutual information was *interaction information* (introduced in [1]), defined for three nodes in terms of the joint and marginal entropies:

$$I(X:Y:Z) = -H(X,Y,Z) + H(X,Y) + H(X,Z) + H(Y,Z) + \tag{1}$$
$$- H(X) - H(Y) - H(Z)\ . \tag{2}$$

Unlike mutual information, this quantity carries a sign. This is traditionally interpreted as the effect that conditioning has on correlation (see [2]):

- $I > 0$: *synergy*. Conditioning on one node *increases* the correlation between the remaining nodes. Example: XOR function.
- $I < 0$: *redundancy*. Conditioning on one node *decreases*, or *explains away* the correlation between the remaining nodes. Example: Ising potential.
- $I = 0$: *3-independence*. Conditioning on one node has no effect on the correlation between the remaining nodes. The nodes can nevertheless still be conditionally dependent. Example: independent nodes.[1]

As argued in [3], [4], and [5], however, the increase or decrease in correlation is not the whole picture. There are systems which exhibit both synergetic and redundant behavior, and interaction information only quantifies the average difference of synergy and redundancy. In a system with highly correlated inputs, for example, the synergy would remain unseen, as it would be cancelled by the redundancy. Moreover, this picture breaks down for more than three nodes. Another problem, pointed out in [5] and [6], is that redundancy (as in the Ising model) can be described in terms of pairwise interactions, not triple, while synergy (as in the XOR function) is purely threewise. Therefore, I compares and mixes information quantities of different nature.

A widely accepted approach, presented in [4] and equivalently in [7], proposed an unsigned measure of synergy. However, it was proven in [8] that such an approach can *not* work in the desired way for more than three nodes.

[1] For an example in which $I = 0$ but the nodes are not independent, see [3].

Technical Definitions

We consider a set of N input nodes $V = \{1, \dots, N\}$, taking values in the sets $X_1, \dots, X_N$, and an output node, taking values in the set Y. We write the input globally as $X := X_1 \times \cdots \times X_N$. For example, in biology Y can be the phenotype, and X can be a collection of genes determining Y. We denote by $F(Y)$ the set of functions on Y, and with $P(X)$ the set of probability measures on X.

We can model the channel from X to Y as a Markov kernel K, that assigns to each $x \in X$ a probability measure on Y (for a detailed treatment, see [15]). Here we will consider only finite systems, so we can think of a Markov kernel simply as a transition matrix, whose rows sum to one.

$$K = K(x, y), \qquad \sum_y K(x, y) = 1 \quad \forall x \, . \tag{3}$$

The pull-back of a function $f : Y \to R$ is:

$$K^* f(x) := \sum_y K(x, y) \, f(y) \, . \tag{4}$$

The push-forward of a probability distribution p on X is:

$$K_* p(y) := \sum_x p(x) \, K(x, y) \, . \tag{5}$$

Let $I \subseteq V$. We would like to restrict our function space (resp. probability space) from V to I, in order to isolate the interactions that take place only within I. In more rigor, given a function $f \in F(X)$, we want to define a particular function $f_I \in F(X)$ which depends only on the I entries (i.e. $f_I \in F(X_I)$.

For brevity, we will denote $F(X_I)$ by F_I. In the extreme cases, $f_V = f$, and $f_\emptyset$ is constant (equal to $\mathbb{E}_p(f)$ for any p). So $F_V = F(X)$, and $F_\emptyset \cong \mathbb{R}$.

Definition. For any such construction, we call:

- $F_\emptyset := F_0$ the space of constant functions, or constant space;
- $F_{\{i\}} := F_i$, with $i = 1, \dots, N$, the single node spaces;
- $F_{\{ij\}} := F_{ij}$, with $1 \leq i < j \leq N$, the 2-interaction spaces;
- $F_{\{i_1, \dots, i_k\}} := F_{i_1, \dots, i_k}$, with $1 \leq i_1 < \cdots < i_k \leq N$ and $0 \leq k \leq N$, the k-interaction spaces.

Let use denote the elements of X_I by x_I. For later convenience, we also introduce the complement:

$$I^c := V \backslash I \, , \tag{6}$$

so that $V = I \cup I^c$, $I \cap I^C = \emptyset$, and $f(x) = f(x_I, x_{I^c})$.

The very same (dual) construction can be made for probability distributions. In this case we have:

- $P_\emptyset := P_0$ contains only the constant measure;
- $P_{\{i\}} := P_i$, with $i = 1, \ldots, N$, contains the marginals on the single node i;
- $P_{\{ij\}} := P_{ij}$, with $1 \leq i < j \leq N$, contains the marginals on nodes i, j;
- $P_{\{i_1,\ldots,i_k\}} := P_{i_1,\ldots,i_k}$, with $1 \leq i_1 < \cdots < i_k \leq N$ and $0 \leq k \leq N$, contains the marginals on the k nodes $i_1, \ldots, i_k$.

1 Linear Decomposition

Let p be a strictly positive probability measure on X. Following the ideas of [9] and [16], we define the following function in F_I:

$$f_I(x) := \sum_{x'_{I^c}} p(x'_{I^c}|x_I)\, f(x_I, x'_{I^c}) \,. \tag{7}$$

First, we have the following consistency results:

Proposition. *With the definition* (7)*, $f_I = f$ if and only if $f \in F_I$.*

Corollary. *Every function of F_I can be written as f_I for some $f \in F(X)$.*

We are saying in other words that the linear map $f \mapsto f_I$, which we denote by Π_I, is idempotent, i.e. a projector. There is more: it is *orthogonal* in the Hilbert space $L^2(X, p)$, whose underlying vector space is $F(X)$, and whose inner product is given by:

$$\langle\, f, g \,\rangle_p := \mathbb{E}_p(fg) = \sum_x p(x)\, f(x)\, g(x) \,. \tag{8}$$

Proposition. *The map $\Pi_I : f \mapsto f_I$ is the* orthogonal projector *onto F_I.*

Sketch of Proof. Because of Proposition 1, the image of Π_I is exactly F_I. We have already seen seen that Π_I is idempotent. By rearranging the sums, one can show that:

$$\langle\, f, \Pi_I g \,\rangle_p = \langle\, \Pi_I f, g \,\rangle_p \,, \tag{9}$$

i.e. Π_I is self-adjoint. □

Remark. For every $I \subseteq V$, $L^2(X_I, p)$ is a Hilbert subspace of $L^2(X, p)$, with underlying space F_I. Moreover, for $I, J \subseteq V$:

- $F_I \cap F_J = F_{I \cap J}$;
- $F_I + F_J \subset F_{I \cup J}$.

The failure of the latter to be an equality is precisely the notion of synergy.

We have a hierarchical structure of linear subspaces:

$$F_\emptyset \subset F_i \subset F_{ij} \subset F_{ijk} \subset \cdots \subset F(X) \ , \tag{10}$$

and projectors:

$$\Pi_\emptyset \leq \Pi_i \leq \Pi_{ij} \leq \Pi_{ijk} \leq \cdots \leq \mathrm{id} \ , \tag{11}$$

where for operators, $A \leq B$ means that $(B-A)$ is positive semidefinite. The projectors have the property that $\Pi_I \Pi_J = \Pi_J = \Pi_J \Pi_I$ if and only if $J \subseteq I$.

We are interested in *pure* interactions. This means that, for $I \subseteq V$, we would like to find the functions f_I that *cannot* be written as (sums of) f_J, with $J \subsetneq I$ strictly. For example, if $I = \{1,2\}$, we are interested in functions of the form $f(x_1, x_2)$, but not of the form $f(x_1), f(x_2)$. Moreover, any function in the form $f(x_1, x_2) + g_1(x_1) + g_2(x_2)$ is also of the type $f(x_1, x_2)$, and we would like somehow to "isolate" the interesting part. In a vector space, this is precisely accomplished by the notion of quotient space. We are interested in the spaces:

$$F_I \Big/ \sum_{J \subsetneq I} F_J \ . \tag{12}$$

Since we are in a Hilbert space, we can work with orthogonal complements instead of equivalence classes. Our quotient space is therefore isomorphic to:

$$F_I \cap \left(\sum_{J \subsetneq I} F_J \right)^\perp := \tilde{F}_I \ . \tag{13}$$

Definition. We call the $\tilde{F}_I$ defined in (13) the *pure I-interaction spaces.*

We denote the orthogonal projector on $\tilde{F}_I$ by $\tilde{\Pi}_I$.

Proposition. *Every I-interaction space is spanned by the pure interaction spaces of lower and equal order:*

$$F_I = \sum_{J \subseteq I} \tilde{F}_J \ . \tag{14}$$

For example, if $I = \{1,2\}$, $F_{12} = \tilde{F}_\emptyset + \tilde{F}_1 + \tilde{F}_2 + \tilde{F}_{12}$.

Remark. Since the pure interaction spaces are all independent, the sums are direct sums:

$$F_I = \bigoplus_{J \subseteq I} \tilde{F}_J \ . \tag{15}$$

This means that given any $f \in F$, we can write it uniquely as a sum:

$$f = \sum_I g_I \ , \tag{16}$$

where each $g_I \in \tilde{F}_I$. Anyway, the g_I in general do *not* correspond to the projections $\tilde{\Pi}_I f$, they only do when the inputs are independent. In the latter case, we have:

$$\Pi_I = \sum_{J \subseteq I} \tilde{\Pi}_J = \bigoplus_{J \subseteq I} \tilde{\Pi}_J \ , \tag{17}$$

and a closed-form expression for the $\tilde{\Pi}_I$ is given by the Moebius inversion theorem (see [16]), which states that (17) is equivalent to:

$$\tilde{\Pi}_I = \sum_{J \subseteq I} (-1)^{\#(I \setminus J)} \Pi_J \ , \tag{18}$$

where $\#(I \setminus J)$ is the number of elements of $I \setminus J$. In general, no such closed form exists for correlated inputs.

Consider now a Markov kernel from X to Y. The projection (7) on X_I implies that if $f : Y \to \mathbb{R}$:

$$(K^* f)_1(x_1) = \sum_{x_2'} p(x_2'|x_1)\, (K^* f)(x_1, x_2') = \sum_{x_2'} p(x_2'|x_1) \sum_y K(x_1, x_2', y)\, f(y) \ .$$

This allows to extend the projections to Markov kernels in the most natural way:

$$K_1^* : f \mapsto K_1^* f := (K^* f)_1 \ . \tag{19}$$

Equivalently, the entries of K_1 are given by:

$$K_1(x_1, y) := \sum_{x_2'} p(x_2'|x_1)\, K(x_1, x_2', y) \ . \tag{20}$$

For binary nodes, the spaces spanned by the K_I are 1-dimensional, so the amount of interaction is determined by the coefficient relative to the only basis element. For higher numbers of states, the projections will be vectors, so it is best to consider their squared norm.

Remark. The projection on a subset I, denoted above by K_I, is a well-defined Markov kernel, i.e. it is positive: if f is a non-negative function, then K_I^* is also non-negative. The projections on *pure* interaction spaces, however, do *not* yield positive linear maps. This means that the objects in the linear decomposition of a Markov kernels are not all themselves Markov kernels.

Examples. Here are some examples of decomposition for binary nodes, with constant input distribution.

- The constant channel is simply a channel that returns 0 or 1 with probability 1/2, regardless of the input.
- The channels x_1 and x_2 copy the respective input, and forget the other one.

- AND, OR, and XOR are the standard Boolean functions.

Channel	K_1	K_2	K_{12}
const	0	0	0
X_1	1	0	0
X_2	0	1	0
AND	1/2	1/2	-1/2
OR	1/2	1/2	1/2
XOR	0	0	1

Channel	$\lvert K_1\rvert^2$	$\lvert K_2\rvert^2$	$\lvert K_{12}\rvert^2$
const	0	0	0
X_1	1	0	0
X_2	0	1	0
AND	1/4	1/4	1/4
OR	1/4	1/4	1/4
XOR	0	0	1

Since the terms in the decomposition are projections, or equivalently (since the spaces here are 1-dimensional) coefficients relative to some basis elements, the terms carry a sign. In the table on the right there are the squared moduli, which can be more useful in higher dimensions.

2 Algebraic Decomposition

The approach of Section 1 was linear, here we have a multiplicative analogue. Or equivalently, we look at a linear decomposition of the *exponent* of the Markov kernel components.

We can see how to decompose the exponent by looking at the expansion:

$$K(x,y) = \frac{1}{Z}\,\exp\left(\sum_I q_I \prod_{i\in I} x_i\, y\right); \qquad Z = \sum_{y'} \exp\left(\sum_I q_I \prod_{i\in I} x_i\, y'\right). \tag{21}$$

The coefficients q_I measure exactly the interaction of the subset I in determining Y. We can see this in the following way. A standard observation in algebraic statistics (see [17]) states that a function $f(x_1, x_2)$ can be written as a product $f_1(x_1)f_2(x_2)$ if and only if:

$$f(x_1,x_2)f(x_1',x_2') = f(x_1',x_2)f(x_1,x_2') \tag{22}$$

for any $x_1' \neq x_1$ and $x_2' \neq x_2$. For n arguments, the function $f(x_1,\dots,x_n)$ can be written as a product of (any) less variable functions if an only if for any $x_i' \neq x_i$:

$$\prod_{\substack{I\subseteq V\\ |I| \text{ even}}} f(x_I', x_{I^C}) = \prod_{\substack{I\subseteq V\\ |I| \text{ odd}}} f(x_I', x_{I^C}). \tag{23}$$

It seems therefore natural to look at the quantity:

$$\frac{p(x_1,x_2)p(x_1',x_2')}{p(x_1',x_2)p(x_1,x_2')} \tag{24}$$

as a natural measure of how far $p(x_1,x_2)$ is from a split probability (see [6], section IV.E). Applying the same idea for Markov kernels, the quantity:

$$\prod_{x_2} \frac{K(x_1,x_2,y)K(x_1',x_2,y')}{K(x_1',x_2,y)K(x_1,x_2,y')}\,, \tag{25}$$

independent of x_2, measures how much y depends only on x_1, and:

$$\frac{K(x_1,x_2,y)K(x_1',x_2',y)K(x_1',x_2,y')K(x_1,x_2',y')}{K(x_1',x_2,y)K(x_1,x_2',y)K(x_1,x_2,y')K(x_1',x_2',y')} \tag{26}$$

measures how much interaction there is between x_1, x_2 in determining y. It is easier to look at the logarithm of such quantities, and so we define for example:

$$q_{ij}(x,x',y,y') := \log \frac{K(x_1,x_2,y)K(x_1',x_2',y)K(x_1',x_2,y')K(x_1,x_2',y')}{K(x_1',x_2,y)K(x_1,x_2',y)K(x_1,x_2,y')K(x_1',x_2',y')} \tag{27}$$

$$= \sum_{J\subseteq\{0,1,2\}} (-1)^{|J|} \log K(x'_J, x_{J^c}) \,, \tag{28}$$

where $X_0 := Y$ for more compact notation. The general formula fo the subset I is (where again $X_0 := Y$):

$$q_I(x,x',y,y') := \sum_{x_{I^c}} \sum_{J\subseteq I\cup\{0\}} (-1)^{|J|} \log K(x'_J, x_{J^c}) \,. \tag{29}$$

It is clear from the definition that these quantities:

- do not depend on the input distribution;
- are *not* defined if some states have zero probability.

Just as in Section 1, for binary nodes, for each node there is only one state x_i' different from x_i, and so the approach is particularly simple. For more than binary inputs, the choice of different $x_i' \neq x_i$ spans a subspace of dimension higher than one. The logarithm of the kernel is itself a function, so we can decompose it exactly like in Section 1. If we take the norm given by the counting measure on the inputs, the projections are *exactly* the q_I, and since the constant measure is split, the alternating sum in (29) corresponds exactly to the Moebius inversion (18). So as in Section 1, for binary nodes we can look at signed coefficients, while for higher dimensions we can look at squared norms.

Examples. Here are some examples of decomposition for binary nodes, with constant input distribution. The channels are the same as for Section 1, but averaged (with weight 1/2) with a constant probability distribution, as deterministic channels cannot be analyzed by this method.

Channel	q_1	q_2	q_{12}
const	0	0	0
X_1	1.6	0	0
X_2	0	1.6	0
AND	0.8	0.8	-0.8
OR	0.8	0.8	0.8
XOR	0	0	1.6

Channel	$\lvert q_1\rvert^2$	$\lvert q_2\rvert^2$	$\lvert q_{12}\rvert^2$
const	0	0	0
X_1	2.4	0	0
X_2	0	2.4	0
AND	0.6	0.6	0.6
OR	0.6	0.6	0.6
XOR	0	0	2.4

As in Section 1, since the q_I are coefficients relative to some basis elements, they are signed. The table on the right gives the squared moduli. Because of the averaging, only relative sizes matter.

3 Geometric Decomposition

Here we take a slightly different approach from the previous two sections. The hierarchical structure is the same, but instead of projections (vectors or signed coefficients), we look at "distances", or in rigor, divergences. Therefore the results will always be non-negative scalars.

In analogy with decompositions (14) and (21), here we define the exponential families:

$$\mathcal{K}_I := \left\{ \frac{1}{Z} \exp\left(\sum_{J\subseteq I} q_J \prod_{i\in J} x_i\, y \right) \middle|\; Z = \sum_{y'} \exp\left(\sum_{J\subseteq I} q_J \prod_{i\in J} x_i\, y' \right), q_J \in \mathbb{R} \right\} . \tag{30}$$

For example, $\mathcal{K}_1$ is the space of (strictly positive) Markov kernels from X to Y which only depend on x_1.

What is the "optimal representative" in $\mathcal{K}_1$ of a given Markov kernel K (the latter possibly depending on x_2)? We want to extend the notion of divergence from probability distributions to Markov kernels. The most natural way of doing it is the following.

Definition. Let p be a strictly positive probability distribution on X, let K, M be strictly positive Markov kernels from X to Y. Then:

$$D_p(K||M) := \sum_{x,y} p(x)\, K(x,y)\, \log \frac{K(x,y)}{M(x,y)} . \tag{31}$$

It is worth noticing that D_p is in general *not* equal to $D(K_*p||M_*p)$. But defined this way, D_p is linear on p, and it is a well-defined divergence. This implies an important compatibility property. Let p, q be joint probability distributions on $X\times Y$, and let D be the KL-divergence. Then:

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D_{p(x)}(p(y|x)||q(y|x)) . \tag{32}$$

Now let $\mathcal{K}$ be a family, and K be a kernel. We define the divergence between K and $\mathcal{K}$ as:

$$D_p(K||\mathcal{K}) := \inf_{M\in\mathcal{K}} D_p(K||M) . \tag{33}$$

First we look at the decomposition in the case of two inputs. For this, we need to define the family of kernels depending on X_1 and X_2, but with no interaction:

$$\tilde{\mathcal{K}}_{12} := \left\{ \frac{1}{Z} \exp\left(q_1x_1y + q_2x_2y\right) \middle|\; Z = \sum_{y'} \exp\left(q_1x_1y' + q_2x_2y'\right), q_1, q_2 \in \mathbb{R} \right\} . \tag{34}$$

Equivalently, this is can be defined implicitly by $q_{12} = 0$ (see Section 2). We then define:

$$G_1 := D_p(K_1||\mathcal{K}_0) \tag{35}$$

$$G_2 := D_p(K_2||\mathcal{K}_0) \tag{36}$$

$$G_{12} := D_p(K_{12}||\,\tilde{\mathcal{K}}_{12}) . \tag{37}$$

The logic is to evaluate the divergence between the KL-projection to I and the family of lower-order interactions. In general, we define:

$$G_I := D_p(K_I || \tilde{\mathcal{K}}_I) \,, \tag{38}$$

where:

$$\tilde{\mathcal{K}}_I := \left\{ \frac{1}{Z} \exp\Big(\sum_{J \subsetneq I} q_J \prod_{i \in J} x_i \, y \Big) \,\Big|\, Z = \sum_{y'} \exp\Big(\sum_{J \subsetneq I} q_J \prod_{i \in J} x_i \, y' \Big), q_J \in \mathbb{R} \right\} . \tag{39}$$

Note the difference with equation (30), here we are taking *proper* subsets of I, neglecting higher interactions (in the language of Section 2, $q_I = 0$).

The projections are easy to compute in the case of single input nodes ($I = \{i\}$), but complicated for higher interaction, for which there is no closed form. The standard approximation algorithm is the iterative scaling (see [11]), which we used to compute the examples below.

It turns out that the projected kernel on a subset via the KL divergence, in the strictly positive case, is equivalent to the linear projection of Section 1.

Proposition. *Let p be a strictly positive probability distribution on X, and let K be a strictly positive Markov kernel from X to Y. Consider the infimum, as in* (33)*:*

$$D_p(K||\mathcal{K}_I) := \inf_{M \in \mathcal{K}_I} D_p(K||M) \,. \tag{40}$$

Then the infimum is realized by K_I, as defined in (19) *and* (20)*.*

Remark. This holds *only* hold for KL-projections on the families $\mathcal{K}_I$, not on other families. For example, it does not hold in general for the $\tilde{\mathcal{K}}_I$. Moreover, this does *not* apply to the deterministic channels (like most examples shown here), since they are not strictly positive.

Examples. Here are some examples of decomposition for binary nodes, with constant input distribution. The channels are again the same as for Section 1.

Channel	G_1	G_2	G_{12}
const	0	0	0
X_1	1	0	0
X_2	0	1	0
AND	0.3	0.3	$1.2 \cdot 10^{-5}$
OR	0.3	0.3	$1.2 \cdot 10^{-5}$
XOR	0	0	1

4 Conclusion

We presented three different ways of decomposing Markov kernels. They all have advantages and disadvantages, sometimes they yield similar or equal results, and in our simple examples they tend to qualitatively agree.

The pros and contras of the different methods can be summarized in the following table. "Y" means "yes", "N" means "no", "I" means "only for independent inputs", and "P" means "only in the strictly positive case".

Property	Linear	Algebraic	Geometric
Well-defined for any input	P	P	Y
Independent on the input	N	P	N[2]
Defined for deterministic channels	Y	N	Y
Zero iff no highest interaction	N	Y	Y
Independent from lower interactions	Y	P	Y
Closed form results	I	Y	N
Computationally simple	I	Y	N

These three approaches can yield improvements over the previous measures of interaction (see the Introduction). In particular:

- They all work for an arbitrary number of nodes;
- They (1 and 3) can tell apart interactions of different orders, without mixing them;
- They (2 and 3) can tell without ambiguity when there is no interaction of a given order or subset.

References

[1] McGill, W. L. *Multivariate information transmission.* Psychometrika, 19(2):97–116, 1954.

[2] Schneidmann, E., Bialek, W., and Berry II, M. J. *Synergy, redundancy, and independence in population codes.* The Journal of Neuroscience, 23(37):11539–11553, 2003.

[3] Williams, P. L. and Beer, R. D. *Nonnegative decomposition of multivariate information.* Preprint available on arXiv:1004.2151, 2010.

[4] Griffith, V. and Koch, C. *Quantifying synergistic mutual information.* Preprint available on arXiv:1205.4265, 2014.

[5] Schneidmann, E., Still, S., Berry II, M. J., and Bialek, W., *Network information and connected correlations.* Physical Review Letters, 91(23):238701-238704, 2003.

[2] How the divergences depend on the input correlation is an interesting open problem. The authors are currently working on it.

[6] Amari, S. *Information geometry on a hierarchy of probability distributions.* IEEE Transactions on information Theory, 47(5):1701–1709, 2001.

[7] Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. *Quantifying unique information.* Entropy, 16, 2014.

[8] Rauh, J., Bertschinger, N., Olbrich, E., and Jost, J. *Reconsidering unique information: towards a multivariate information decomposition.* Preprint available on arXiv:1404.3146, 2015.

[9] Darroch, J. N. and Speed, T. P. *Additive and multiplicative models and interactions.* Annals of Statistics, 11:724–738, 1983.

[10] Amari, S. and Nagaoka, H. *Differential geometry of smooth families of probability distributions.* Preprint available on Technical Report METR 82-7, Univ. of Tokyo, 1982.

[11] Csiszár, I. and Shields, P. C. *Information Theory and Statistics: A Tutorial.* Foundations and Trends in Communications and Information Theory, 1(4):417–528, 2004

[12] Ay, N. and Knauf, A. *Maximizing Multi-Information.* Kybernetica, 42, 2006

[13] Ay, N. *An Information-Geometric Approach to a Theory of Pragmatic Structuring.* The Annals of Probability, 30, 2002

[14] Nagaoka, H. *The exponential Family of Markov Chains and Its Information Geometry.* 28th Symposium on Information Theory and Applications, 2005

[15] Kakihara, Y. *Abstract Methods in Information Theory.* World Scientific, 1999.

[16] Lauritzen, S. L. *Graphical Models.* Oxford, 1996.

[17] Drton, M., Sturmfels, B., and Sullivant, S. *Lectures on Algebraic Statistics.* Birkhaeuser, 2009.

[18] Amari, S. and Nagaoka, H. *Methods of Information Geometry.* Oxford, 1993.

A New Method for tackling Asymmetric Decision Problems

Peter A. Thwaites
School of Mathematics
University of Leeds
P.A.Thwaites@leeds.ac.uk

Jim Q. Smith
Department of Statistics
University of Warwick
J.Q.Smith@warwick.ac.uk

Abstract

Chain Event Graphs are probabilistic graphical models designed especially for the analysis of discrete statistical problems which do not admit a natural product space structure. We show here how they can be used for decision analysis, and describe an optimal decision strategy based on an efficient local computation message passing scheme. We briefly describe a method for producing a parsimonious decision CEG, analogous to the parsimonious ID, and touch upon the CEG-analogues of Shachter's barren node deletion and arc reversal for ID-based solution.
Keywords: Chain Event Graph, decision analysis, Influence diagram

1 Introduction

In this paper we demonstrate how the Chain Event Graph (CEG) (see for example [14, 16, 15, 12, 1]) can be used for tackling asymmetric decision problems.

Extensive form (EF) decision trees (in which variables appear in the order in which they are observed by a decision maker) are flexible and expressive enough to represent asymmetries within both the decision and outcome spaces, doing this through the topological structure of the tree. They can however become unwieldy, and are not convenient representations from which to read the conditional independence structure of a problem.

Other graphical representations have been developed which to some extent deal with the complexity issue associated with decision trees, and also allow for local computation. The most commonly used of these is the Influence diagram (ID). Because of their popularity, ID solution techniques have developed considerably since their first introduction. However a major drawback of the ID representation is that many decision problems are asymmetric in that different actions can result in different choices in the future, and IDs are not ideally suited to this sort of problem [5]. As decision makers have become more ambitious in the complexity of the problems they wish to solve, standard ID and tree-based methods have proven to be inadequate, and new techniques have become necessary.

There have consequently been many attempts to adapt IDs for use with asymmetric problems (see for example [13, 9, 8]), or to develop new techniques which use both IDs and trees [4]. There have also been several new structures suggested, such as Sequential Decision Diagrams (SDDs) [5] and Valuation Networks (VNs) [11]. Asymmetric problems have recently also been represented via decision circuits [2]. An overview of many of these developments is given by Bielza & Shenoy in [3]. They note that none of the methods available is consistently better than the others.

CEGs are probabilistic graphical models designed especially for the representation and analysis of discrete statistical problems which do not admit a natural product space structure. Unlike Bayesian Networks (BNs) they are functions of event trees, and this means that they are able to express the complete sample space structure associated with a problem. They are particularly useful for the analysis of processes where the future development at any specific point depends on the particular history of the problem up to that point. Such dependencies can be thought of as context-specific conditional independence properties; and the structure implied by these properties is fully expressed by the topology of the CEG. This is a distinct advantage over context-specific BNs, which require supplementary information usually in the form of trees or conditional probability tables attached to some of the vertices of the graph. Like BNs, CEGs provide a suitable framework for efficient local computation algorithms.

Using CEGs for asymmetric decision analysis overcomes drawbacks associated with the current graphs and techniques used for this purpose. They are an advance on decision trees as they encode the conditional independence structure of problems. They can represent probability models consistently (which SDDs don't), and do not require dummy states or supplementing with extra tables or trees (a drawback of both VNs and Smith et als' adaptations of IDs). They can model all asymmetries (which VNs cannot), and their semantics are very straightforward, making them an appropriate tool for use by non-experts (both VN & SDD methodologies are very complicated).

Call & Miller [4] have drawn attention to the value of coalescence in tree-based approaches to decision problems. They also point out that the difficulties in reading conditional independence structure from trees has meant that analysts using them have not fully taken advantage of the idea of coalescence. They remark that *the ability to exploit asymmetry can be a substantial advantage for trees. If trees could naturally exploit coalescence, the efficiency advantage is even greater.* SDDs go some way towards exploiting this [3], but decision CEGs use coalescence both as a key tool for the expression of conditional independence structure, and to power the analysis.

We show here how CEGs can be used for decision analysis, and describe how to arrive at an optimal decision strategy via an efficient local computation message passing scheme. We briefly describe a method for producing a parsimonious decision CEG, analogous to the parsimonious ID, which contains only those variables and dependencies which the decision maker needs to consider when making decisions; and touch upon the CEG-analogues of Shachter's [10] barren node deletion and arc reversal for ID-based solution.

2 CEGs and decision CEGs

We start this section with a brief introduction to CEGs – we direct readers to one of [14, 15] if they would like a more detailed definition. The CEG is a function of a coloured event tree, so we begin with a description of these graphs.

- A coloured event tree $\mathcal{T}$ is a directed tree with a single root-node.
- Each non-leaf-node v has an associated random variable whose state space corresponds to the subset of directed edges of $\mathcal{T}$ which emanate from v.
- Each edge leaving a node v carries a *label* which identifies a possible immediate future development given the partial history corresponding to the node v.
- The non-leaf-node set of $\mathcal{T}$ is partitioned into equivalence classes called *stages*: Nodes in the same *stage* have sets of outgoing edges with the same labels, and edges with the same labels have the same probabilities.
- The edge-set of $\mathcal{T}$ is partitioned into equivalence classes, whose members have the same *colour*: Edges have the same *colour* when the vertices from which they emanate are in the same stage and the edges have the same label (& hence probability).
- The non-leaf-node set of $\mathcal{T}$ is also partitioned into equivalence classes called *positions*: Nodes are in the same *position* if the *coloured subtrees* rooted in these nodes are isomorphic both in topology and in colouring (so edges in one subtree are coloured (and labelled) identically with their corresponding edges in another).

Note that nodes are in the same position when the sets of complete future developments from each node are the same, and have the same probability distribution.

To produce a CEG $\mathcal{C}$ from our tree $\mathcal{T}$, nodes in the same position are combined (as in the coalesced tree), and all leaf-nodes are combined into a single sink-node. We note that for CEGs used for decision problems it is often more convenient to replace the single sink-node by a set of terminal utility nodes, each of which corresponds to a different utility value. We return to this idea in our example in Section 3.

So the nodes of our CEG $\mathcal{C}$ are the *positions* of the underlying tree $\mathcal{T}$. We transfer the ideas of *stage* and *colour* from $\mathcal{T}$ to $\mathcal{C}$, and it is this combination of positions and stages that enables the CEG to encode the full conditional independence structure of the problem being modelled [14].

Many discrete statistical processes are asymmetric in that some variables have quite different collections of possible outcomes given different developments of the process up to that point. It was for these sorts of problem that the CEG was created, and one area where they have proved particularly useful is that of causal analysis [16, 15]. In much causal analysis the question being asked is *If I make this manipulation, what are the effects?*, but graphical models set up to answer such questions can also be readily used for questions such as *If I want to maximise my utility over this process, what are the manipulations (decisions) I need to make?*

In attempting to answer this second question, we notice that there are only certain nodes or positions in the CEG which can actually be manipulated. We concentrate in this paper on manipulations which impose a probability of one onto one edge

emanating from any such node (equivalent to making a firm decision). Hence the probabilistic nature of these nodes is removed – they become decision nodes, and we therefore draw them as squares.

We draw our CEG in EF order – as with decision trees this is necessary in order to calculate optimal decision rules. If two decision nodes in $\mathcal{T}$ are in the same position, then the optimal strategy is the same for the decision maker (DM) at each of the two decision nodes: it is conditionally independent of the path taken to reach the decision node. A similar interpretation can be given to two chance nodes in the same position.

The only other modification that is required to use the CEG for decision analysis is the addition of utilities. This can be done in two ways (1) adding utilities to edges, or (2) expanding the sink-node w_∞ into a set of utility nodes, each corresponding to a distinct utility value (see our example in Section 3). We make our terminal nodes diamond-shaped whether they are leaf nodes or a single sink-node.

When we manipulate a CEG we prune edges that are given zero probability by the manipulation, and also any edge or position which lies downstream of such edges only. No other edges (except those we manipulate to) have probabilities changed by the manipulation [15]. This is not the case when we simply observe an event, when edge-probabilities upstream of the observation can also change.

In [7], Dawid outlines how a decision-theoretic approach can be taken to causal inference. In this paper we are perhaps doing the opposite; we show how established causal analysis techniques for CEGs have a natural application in the field of decision analysis.

Our propagation algorithm is illustrated in Table 1 – at the end of the local message passing, the root node will contain the maximum expected utility. In the pseudocode we use C & D for the sets of chance & decision nodes, p represents a probability or weight, and u a utility. The utility part of a position w is denoted by $w[u]$, the probability part of an edge by $e(w, w')[p]$ etc. The set of child nodes of a position w is denoted by $ch(w)$. Note that there may be more than one edge connecting two positions, if say two different decisions have the same consequence. This has significant ramifications for more complicated problems, as described in our example.

Table 1: Local propagation algorithm for finding an optimal decision sequence

- Find a topological ordering of the positions. Without loss of generality call this $w_1, w_2, \ldots, w_n$, so that w_1 is the root-node, and w_n is the sink-node.
- Initialize the utility value $w_n[u]$ of the sink node to zero.
- Iterate: for $i = n - 1$ step minus 1 until $i = 1$ do:
  - If $w_i \in C$ then $w_i[u] = \sum_{w \in ch(w_i)} \left[\sum_{e(w_i,w)} \left[e(w_i, w)[p] * (w[u] + e(w_i, w)[u]) \right] \right]$
  - If $w_i \in D$ then $w_i[u] = \max_{w \in ch(w_i)} \left[\max_{e(w_i,w)} \left[(w[u] + e(w_i, w)[u]) \right] \right]$
- Mark the sub-optimal edges.

Note that when we choose to confine utilities to terminal utility nodes, this algorithm is much simplified since both the initializing step and the $e(w_i, w)[u]$ compo-

nents are no longer required.

3 Representing and solving asymmetric decision problems using extensive form CEGs

We concentrate here on how the CEG compares with the augmented ID of Smith, Holtzman & Matheson [13] for the representation and solution of asymmetric decision problems. We show that the ID-based solution techniques of barren-node deletion [10] and parsimony have direct analogues in the CEG-analysis, and that arc-reversal [10] is not required for the solution of EF CEGs. The *distribution trees* [13] added to the nodes of IDs to describe the asymmetry of a problem can simply be thought of as close-ups of interesting parts of the CEG-depiction, where they are an integral part of the representation rather than bolt-on as is the case with IDs. We illustrate this comparison through an example.

We first consider what is meant by conditional independence statements which involve decision variables.

The statement $X \amalg Y \mid Z$ is true if and only if we can write $P(x \mid y, z)$ as $a(x, z)$ for some function a of x and z, for all values x, y, z of the variables X, Y, Z [6]. So clearly, for chance variables X, Y, Z and decision variable D, where the value taken by X is not known to the DM when she makes a decision at D, we can write statements such as $X \amalg D \mid Z$ and $X \amalg Y \mid D$ since the expressions $P(x \mid d, z) = a(x, z)$ and $P(x \mid y, d) = a(x, d)$ are unambiguous in this situation (d representing a value taken by D).

Note that $P(d \mid y, z)$ is not unambiguously defined, and so conditional independence is no longer a symmetric property when we add decision variables to the mix. By a slight abuse of notation we can also write $U \amalg (Y, D_1) \mid (Z, D_2)$ if $U(y, z, d_1, d_2) = U(z, d_2)$ for all values y, z, d_1, d_2 of the chance variables Y, Z and decision variables D_1, D_2.

Example. *Patients suffering from some disease are given one of a set of possible treatments. There is an initial reaction to the treatment in that the patient's body either accepts the treatment without problems or attempts to reject it. After this initial reaction, the patient responds to the treatment at some level measurable by their doctor, and this response is independent of the initial reaction conditioned on which treatment has been given. The patient's doctor has to make a second decision on how to continue treatment.*

There is also the possibility of the patient having some additional condition which affects how they will respond to the treatment. Whether or not they have this condition will remain unknown to the doctor, but she can estimate the probability of a patient having it or not (conditioned on their response to their particular treatment) from previous studies.

The doctor is concerned with the medium-term health of the patient following her decisions, and knows that this is dependent on whether or not the patient has the additional condition, how they respond to the first treatment, and the decision made regarding treatment continuation.

Table 2 summarises this information in the form of a list of variables and relationships.

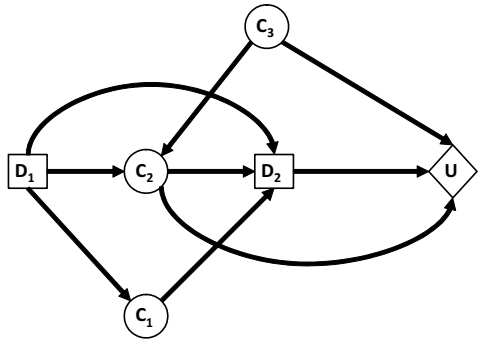


Figure 1: EF ID for our example

Table 2: Variables & relationships; plus U as a function of C_3, D_2 and C_2

		C_3	C_2	D_2	U
D_1:	Choice of treatment	1	1	1	A
C_1:	Initial reaction	1	1	2	A
C_2:	Response to treatment – $C_2 \amalg C_1 \mid D_1$	1	2	1	B
D_2:	Decision on how to continue treatment	1	2	2	C
C_3:	Condition affecting response to	2	1	1	A
	treatment and medium-term health	2	1	2	A
	Can estimate $P(C_3 \mid D_1, C_2)$	2	2	1	D
U:	Medium-term health, a function of	2	2	2	E
	C_2, D_2 and C_3				

To avoid making the problem too complex for easy understanding we let all variables be binary except U, and introduce only two asymmetric features: So suppose that if a patient fails to respond to the first treatment ($C_2 = 1$), then the patient will inevitably have the lowest medium-term health rating ($U = A$). We can express this as $U \amalg (C_3, D_2) \mid (C_2 = 1)$ (see Table 2). Suppose also that if $D_1 = 2$ (Treatment 2 is given) then C_1 takes the value 1 (the patient's body always accepts the treatment). The problem can be represented by the EF ID in Figure 1.

To express the asymmetry of the problem we can add *distribution trees* to the nodes C_1 and U as in Figure 2. These have been drawn in a manner consistent with the other diagrams in this paper, rather than with those in [13].

The ID in Figure 1 is not the most parsimonious representation of the problem. If we can partition the parents of a decision node D (those nodes with arrows into

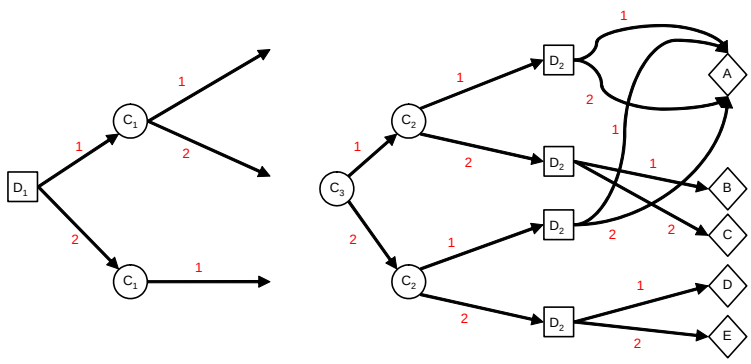


Figure 2: Distribution trees for nodes C_1 and U

D) into two sets $Q^A(D), Q^B(D)$ such that $U \amalg Q^B(D) \mid (D, Q^A(D))$, then the set $Q^B(D)$ can be considered *irrelevant* for the purposes of maximising utility, and the edges from nodes in $Q^B(D)$ into D can be removed from the ID. Here we find that $C_1 \in Q^B(D_2)$, and so the edge from C_1 to D_2 can be removed from the ID. The node C_1 is now barren, so it can also be removed (together with the edge $D_1 \to C_1$).

Once we have our parsimonious ID we can use one of the standard solution methods to produce an optimal decision strategy and expected utility for this strategy. Using Shachter's method (reversing the arc between C_3 and C_2, and adding a new arc from D_1 to C_3) we eventually get

$$U^{final} = \max_{D_1} \Big[\sum_{C_2} P(C_2 \mid D_1) \Big[\max_{D_2} \Big[\sum_{C_3} P(C_3 \mid D_1, C_2)\ U(C_2, C_3, D_2) \Big] \Big] \Big]$$

which does not however reflect the asymmetries in the problem. These can be built into the solution technique, but as the principal asymmetry concerns $U(C_2, C_3, D_2)$, any advantage conveyed by the compactness of the ID is lost in the messy arithmetic.

We now turn our attention to a CEG-representation of the problem. There are two EF orderings of the variables: $D_1, C_2, C_1, D_2, C_3, U$ and one where C_1 & C_2 are interchanged. Note that D_2 precedes C_3 since the value of C_3 is not known to the DM when she comes to make a decision at D_2. The first ordering leads to a slightly more transparent graph.

As we are comparing CEGs and IDs here, we do not put any utilities onto edges, but restrict them to terminal utility nodes. We also separate out our single utility node into distinct utility nodes for each value taken by U. In more complex decision problems this can lead to greater transparency. We have elsewhere called this form of CEG without utilities on edges, and with separated utility nodes, a Type 2 decision CEG. The Type 2 CEG for the ordering $D_1, C_2, C_1, D_2, C_3, U$ is given in Figure 3.

Conditional independence structure in a CEG can be read from individual positions, from stages, and from *cuts* through these [14]. Recall that nodes in the

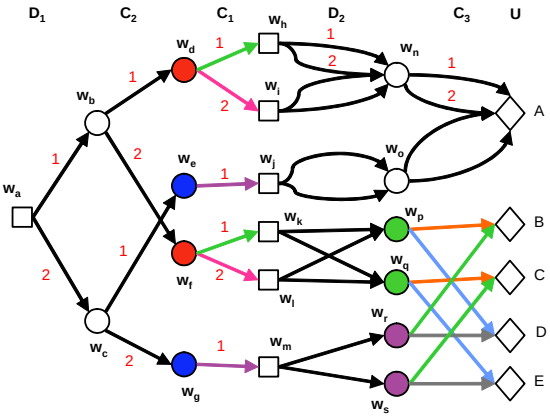


Figure 3: Initial EF CEG for our example

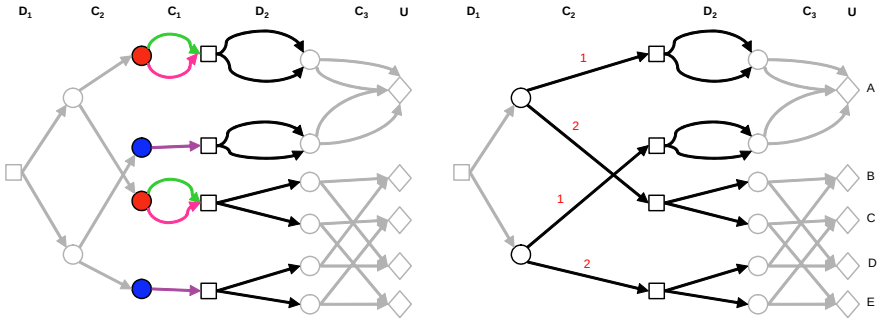


Figure 4: First and second simplifications

underlying tree are coalesced into positions when the sets of complete future developments from each node are the same and have the same probability distribution. So for example, the position w_n yields the information that

$$(C_3, U) \amalg (C_1, D_2) \mid (D_1 = 1, C_2 = 1) \qquad (3.1)$$

The position w_p similarly yields $(C_3, U) \amalg C_1 \mid (D_1 = 1, C_2 = 2, D_2 = 1)$.

Recall that positions in a CEG are in the same stage if their sets of outgoing edges carry the same labels and have the same probability distribution. The positions w_p & w_q are in the same stage (indicated by the colouring), and so the probabilities on the edges leaving these positions have the same distibution, and hence

$$C_3 \amalg (C_1, D_2) \mid (D_1 = 1, C_2 = 2) \qquad (3.2)$$

The expressions (3.1) & (3.2) result from the fact that in our EF CEG ordered $D_1, C_2, C_1, D_2, C_3, U$, the variable C_3 is dependent on D_1 and C_2. This is not clear

from the ID in Figure 1, but is reflected in the expression for U^{final}. But the form of this expected utility expression is a consequence of the arc-reversal required for successful ID-based solution of our problem. So this arc-reversal is already explicitly represented in the original EF CEG, and is not (as with IDs) an additional requirement of the solution technique.

A *cut* through a CEG is a set of positions or stages which partitions the set of root-to-sink/leaf paths. So the set of positions $\{w_n, w_o, w_p, w_q, w_r, w_s\}$ is a cut of our CEG. A conditional independence statement associated with a cut is the union of those statements associated with the component positions (or stages) of the cut. So the cut through $\{w_n, w_o, w_p, w_q, w_r, w_s\}$ gives us that

$$U \amalg C_1 \mid (D_1, C_2, D_2)$$

which is clearly of the form $U \amalg Q(D_2^B) \mid (D_2, Q(D_2^A))$, and tells us that C_1 is irrelevant to D_2 for the purposes of maximising utility.

For a Type 2 CEG drawn in EF order, two (or more) decision nodes are in the same position if the sub-CEGs rooted in each decision node have the same topology, equivalent edges in these sub-CEGs have the same labels & (where appropriate) probabilities, and equivalent branches terminate in the same utility node. So in Figure 3, the nodes w_h & w_i are in the same position, as are the nodes w_k & w_l. Decision nodes in the same position can simply be coalesced, giving us the first graph in Figure 4.

For a Type 2 EF decision CEG with all positions coalesced (as in this graph), a barren node is simply a position w for which $ch(w)$ (defined as in section 2) contains a single element. Barren nodes can be deleted in a similar manner to those in BNs – see Table 3 (where $pa(w)$ denotes the set of parent nodes of w).

Table 3: Barren node deletion algorithm (Type 2 decision CEGs)

- Choose a topological ordering of the positions excluding the terminal utility nodes: $w_1, w_2, \ldots, w_m$, such that w_1 is the root-node.
- Iterate: for $i = 2$ step plus 1 until $i = m$ do:
  - If $ch(w_i)$ contains only one node then
    Label this node $w_{\succ i}$
    For each node $w_{\prec i} \in pa(w_i)$
    Replace all edges $e(w_{\prec i}, w_i)$ by a single edge $e(w_{\prec i}, w_{\succ i})$
    Delete all edges $e(w_i, w_{\succ i})$ & the node w_i.

Four iterations of the algorithm applied to the first graph in Figure 4 yield the second graph in Figure 4. Further iterations will remove the first two D_2 nodes and the first two C_3 nodes to give the parsimonious CEG in Figure 5.

We can clearly see that C_1 is irrelevant for maximising U, and moreover if $C_2 = 1$ then both D_2 and C_3 are also irrelevant for this purpose (so the DM actually only needs to make one decision in this context). This latter property of the problem is not one that can be deduced from an ID-representation, although it could with some effort be worked out from the second distribution tree in Figure 2. It is however obvious in the parsimonious CEG.

Solution follows the method described in section 2 (the process obviously being simpler as there are no rewards or costs on the edges), and results in the expression

$$
\begin{aligned}
&U^{final} = \\
&\max\big[P(C_2 = 1 \mid D_1 = 1)U_A + P(C_2 = 2 \mid D_1 = 1)\times \\
&\max\big[P(C_3 = 1 \mid D_1 = 1, C_2 = 2)U_B + P(C_3 = 2 \mid D_1 = 1, C_2 = 2)U_D, \\
&\qquad P(C_3 = 1 \mid D_1 = 1, C_2 = 2)U_C + P(C_3 = 2 \mid D_1 = 1, C_2 = 2)U_E\big], \\
&\qquad P(C_2 = 1 \mid D_1 = 2)U_A + P(C_2 = 2 \mid D_1 = 2)\times \\
&\max\big[P(C_3 = 1 \mid D_1 = 2, C_2 = 2)U_B + P(C_3 = 2 \mid D_1 = 2, C_2 = 2)U_D, \\
&\qquad P(C_3 = 1 \mid D_1 = 2, C_2 = 2)U_C + P(C_3 = 2 \mid D_1 = 2, C_2 = 2)U_E\big]\big]
\end{aligned}
$$

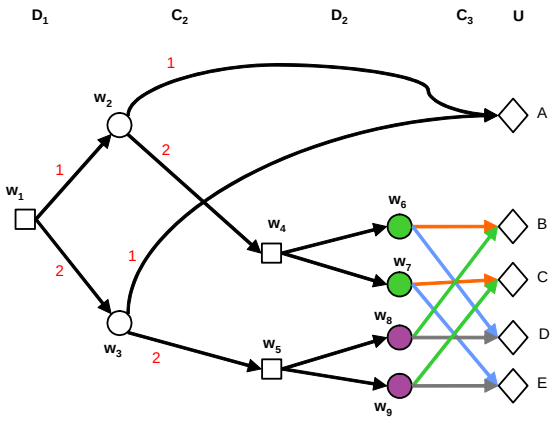


Figure 5: Parsimonious CEG

This expression is obviously more complex than that for the ID, but it is much more robust since it has been produced using the asymmetry of the problem to power the analysis, rather than treating it as an added complication.

4 Discussion

In this paper we have concentrated on how CEGs compare with IDs for the analysis of asymmetric decision problems. It is however worth pointing out two advantages of CEGs over coalesced trees: Firstly, the ability to read conditional independence structure from CEGs allowed us to create an analogue of the parsimonious ID, and secondly, the explicit representation of stage structure in CEGs gave rise to our barren node deletion algorithm.

A paper providing a more detailed discussion of parsimony, barren node deletion and arc reversal as they relate to CEGs is imminent. This paper will also provide a

comparison of CEGs with VNs, SDDs and augmented IDs through a worked example. A further paper on the use of decision CEGs for multi-agent problems and games is also in the pipeline.

Acknowledgement: This research is being supported by the EPSRC (project EP/M018687/1).

References

[1] Barclay L. M., Hutton J. L., and Smith J. Q. (2013) Refining a Bayesian Network using a Chain Event Graph, *International Journal of Approximate Reasoning*, 54:1300–1309.

[2] Bhattacharjya D. and Shachter R. D. (2012) Formulating asymmetric decision problems as decision circuits, *Decision Analysis*, 9:138–145.

[3] Bielza C. and Shenoy P. P. (1999) A comparison of graphical techniques for asymmetric decision problems, *Management Science*, 45:1552–1569.

[4] Call H. J. and Miller W. A. (1990) A comparison of approaches and implementations for automating Decision analysis, *Reliability Engineering and System Safety*, 30:115–162.

[5] Covaliu Z. and Oliver R. M. (1995) Representation and solution of decision problems using sequential decision diagrams, *Management Science*, 41(12).

[6] Dawid A. P. (1979) Conditional independence in statistical theory, *Journal of the Royal Statistical Society, Series B*, 41:1–31.

[7] Dawid A. P. (2012) The decision-theoretic approach to causal inference. In C. Berzuini, A. P. Dawid, and L. Bernardinelli, editors, *Causality: Statistical Perspectives and Applications*, pages 25–42. Wiley.

[8] Jensen F. V., Nielsen T. D., and Shenoy P. P. (2006) Sequential influence diagrams: A unified asymmetry framework, *International Journal of Approximate Reasoning*, 42:101–118.

[9] Qi R., Zhang N., and Poole D. (1994) Solving asymmetric decision problems with influence diagrams. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 491–499.

[10] Shachter R. D. (1986) Evaluating Influence diagrams, *Operations Research*, 34(6):871–882.

[11] Shenoy P. P. (1996) Representing and solving asymmetric decision problems using valuation networks. In D. Fisher and H-J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*. Springer-Verlag.

[12] Silander T. and Leong T-Y. (2013) A Dynamic Programming Algorithm for Learning Chain Event Graphs. In *Discovery Science*, volume 8140 of *Lecture Notes in Computer Science*, pages 201–216. Springer.

[13] Smith J. E., Holtzman S., and Matheson J. E. (1993) Structuring conditional relationships in influence diagrams, *Operations Research*, 41:280–297.

[14] Smith J. Q. and Anderson P. E. (2008) Conditional independence and Chain Event Graphs, *Artificial Intelligence*, 172:42–68.

[15] Thwaites P. A. (2013) Causal identifiability via Chain Event Graphs, *Artificial Intelligence*, 195:291–315.

[16] Thwaites P. A., Smith J. Q., and Riccomagno E. M. (2010) Causal analysis with Chain Event Graphs, *Artificial Intelligence*, 174:889–909.

Relationship of Compositional Models and Networks in Imprecise Probabilities Frameworks

Jiřina Vejnarová
Department of Decision-Making Theory
Institute of Information Theory and Automation of the CAS
vejnar@utia.cas.cz

Abstract

The contribution is devoted to the relationship between a special type of compositional models, so-called prefect sequences, and networks in four particular frameworks of imprecise probabilities. We show that although the class of perfect sequences of probability distributions is equivalent with the class of Bayesian networks (and analogous equivalence holds also in possibilistic setting), the class of evidential compositional models is much wider than that of evidential networks and the relationship among credal networks (in general sense), perfect sequences of credal sets and separately specified credal networks is even more interesting.

Keywords: Compositional models, Bayesian networks, possibilistic networks, evidential networks, credal networks

1 Introduction

Compositional models of precise probability distributions were introduced almost twenty years ago [9] with the aim to bring an alternative to Graphical Markov models. Later the compositional models were introduced also in possibility theory [14] utilizing the formal similarity of possibility and probability theories, more precisely, the ability to express both probability and possibility measures by a point function — distribution. Nevertheless, there exist one substantial difference with probabilistic framework — multidimensional models are parameterized by a continuous t-norm. The generalization of compositional models to evidence theory [11] was not too simple, as it is necessary to work with set functions instead of point ones, and the generalization to even more general framework of credal sets is under development [17].

Bayesian networks, on the other hand, are at present probably the most popular representative of Graphical Markov models. Therefore, it is not surprising that their

counterparts in imprecise probabilities framework have been introduced during last two decades [1, 3, 5].

This contribution is devoted to the overview of compositional models and corresponding networks in four above-mentioned frameworks. We will show that while the class of perfect sequences of probability distributions is equivalent with the class of Bayesian networks (and analogous equivalence holds also in possibilistic setting), the class of evidential compositional models is much wider than that of evidential networks and the relationship among credal networks (in general sense), perfect sequences of credal sets and separately specified credal networks is even more interesting.

The contribution is organised as follows. After an overview of basic concepts in particular frameworks (Section 2), in Section 3 compositional models (and their properties) will be recalled and Section 4 is the overview of relationships between compositional models and networks in particular frameworks.

2 Basic concepts and notation

In this section we will recall basic concepts and notation necessary for understanding the contribution.

2.1 Set projections and extensions

For an index set $N = \{1, 2, \ldots, n\}$ let $\{X_i\}_{i \in N}$ be a system of variables, each X_i having its values in a finite set $\mathbf{X}_i$. In this contribution we will deal with a *multidimensional frame of discernment* (or simply Cartesian product space)

$$\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and, for $K \subseteq N$, its *subframes* (or subspaces)

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

When dealing with groups of variables on these subframes, X_K will denote a group of variables $\{X_i\}_{i \in K}$ throughout the contribution.

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 $K = \{i_1, i_2, \ldots, i_k\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathbf{X}_K.$$

Analogously, for $M \subset K \subseteq N$ and $A \subset \mathbf{X}_K$, $A^{\downarrow M}$ will denote a *projection* of A into $\mathbf{X}_M$. In this case

$$A^{\downarrow M} = \{y \in \mathbf{X}_M \mid \exists x \in A : y = x^{\downarrow M}\}.$$

In addition to the projection, in this text we will also need an inverse operation, which will be called a join [2]. By a *join* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ $(K, L \subseteq N)$, we will understand a set

$$A \bowtie B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \ \ \& \ \ x^{\downarrow L} \in B\}.$$

Let us note that, for any $C \subseteq \mathbf{X}_{K\cup L}$, naturally $C \subseteq C^{\downarrow K} \bowtie C^{\downarrow L}$, but generally $C \neq C^{\downarrow K} \bowtie C^{\downarrow L}$. Furthermore, if K and L are disjoint, then the join of A and B is just their Cartesian product, $A \bowtie B = A \times B$, and if $K = L$ then $A \bowtie B = A \cap B$. If $K \cap L \neq \emptyset$ and $A^{\downarrow K\cap L} \cap B^{\downarrow K\cap L} = \emptyset$ then $A \bowtie B = \emptyset$ as well.

2.2 Probability and possibility distributions

The uncertainty of a group of variables X_K can "traditionally" be described by a *probability distribution* (sometimes also called *probability function*) $P : \mathbf{X}_K \longrightarrow [0, 1]$, such that

$$\sum_{x_K\in\mathbf{X}_K} P(x_K) = 1.$$

Having two probability distributions P_1 and P_2 of X_K we say that P_1 is *absolutely continuous* with respect to P_2 (and denote $P_1 \ll P_2$) if for any $x_K \in \mathbf{X}_K$

$$P_2(x_K) = 0 \Longrightarrow P_1(x_K) = 0.$$

This concept plays an important role in the definition of the probabilistic composition operator.

More specific (but not less important) is the concept of projectivity. Probability distributions P_1 and P_2 of X_K and X_L, respectively, are called *projective*, if they coincide on common subspaces, i.e. if

$$P_1(x_{K\cap L}) = P_2(x_{K\cap L})$$

for any $x_{K\cap L} \in \mathbf{X}_{K\cap L}$.

As an alternative to probability one can use a possibility distribution

$$\pi : \mathbf{X}_K \longrightarrow [0, 1],$$

which is called *normal* if

$$\max_{x_K\in\mathbf{X}_K} \pi(x_K) = 1.$$

In a way closely connected with the notion of normalization is also the most important difference between the two considered settings, which concerns marginalization.

Marginalization in possibility theory differs from that in the probabilistic framework in using maximization instead of summation, i.e. for $L \subset K$ a *marginal possibility distribution* $\pi(x_L)$ of distribution $\pi(x_K)$ is defined by the formula

$$\pi(x_L) = \max_{x_{K\setminus L}\in\mathbf{X}_{K\setminus L}} \pi(x_K) = \max_{x_{K\setminus L}\in\mathbf{X}_{K\setminus L}} \pi(x_L, x_{K\setminus L}).$$

Analogous to probabilistic framework, we say that possibility distributions π_1 and π_2 of X_K and X_L, respectively, are called *projective*, if

$$\pi_1(x_{K\cap L}) = \pi_2(x_{K\cap L})$$

for any $x_{K\cap L} \in \mathbf{X}_{K\cap L}$.

From the point of view of this contribution, one of the most important notions we have to recall is the concept of conditioning [6]. Considering a continuous t-norm[1] T, the *conditional possibility distribution* $\pi\left(x_2|_T x_1\right)$ is defined for each $(x_1, x_2) \in \mathbf{X}_1 \times \mathbf{X}_2$ as *any* solution of the equation

$$\pi\left(x_1, x_2\right) = T\left(\pi\left(x_2|_T x_1\right), \pi\left(x_1\right)\right).$$

Since the solution of this equation is usually not unique, we take for the conditional distribution the maximal (or the least specific) one. As we consider only continuous t-norms T, this solution coincides with the respective *T-residual*

$$y \triangle_T x = \sup\{z \in [0,1] : T(z,x) \leq y\},$$

i.e. for each $(x_1, x_2) \in \mathbf{X}_1 \times \mathbf{X}_2$

$$\pi\left(x_2|_T x_1\right) = \pi\left(x_1, x_2\right) \triangle_T \pi\left(x_1\right).$$

2.3 Set functions

In evidence theory [13], which can be considered as a generalization of both probability and possibility theories, two dual measures are used to model the uncertainty: belief and plausibility measures. Each of them can be defined with the help of another set function called a *basic assignment* m on $\mathbf{X}_N$, i.e.,

$$m : \mathcal{P}(\mathbf{X}_N) \longrightarrow [0,1],$$

where $\mathcal{P}(\mathbf{X}_N)$ is the power set of $\mathbf{X}_N$, and

$$\sum_{A \subseteq \mathbf{X}_N} m(A) = 1.$$

Furthermore, we assume that $m(\emptyset) = 0$.[2] A set $A \in \mathcal{P}(\mathbf{X}_N)$ is a *focal element* if $m(A) > 0$.

For a basic assignment m on $\mathbf{X}_K$ and $M \subset K$, a *marginal basic assignment* of m on $\mathbf{X}_M$ is defined (for each $A \subseteq \mathbf{X}_M$) by the equality

$$m^{\downarrow M}(A) = \sum_{\substack{B \subseteq \mathbf{X}_K \\ B^{\downarrow M} = A}} m(B).$$

Analogous to previous subsection, we will call two basic assignments m_1 and m_2 on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively, *projective*, if

$$m_1^{\downarrow K \cap L}(A) = m_2^{\downarrow K \cap L}(A)$$

for any $A \subseteq \mathbf{X}_{K \cap L}$.

Although there exist a great number of conditioning rules [8], their usefulness for multidimensional models is rather questionable. This fact led us to the following proposal of a new conditioning rule in [16], where also its correctness was proven.

[1] Let us recall that a t-norm T is a commutative, associative and isotone binary operator on [0,1] satisfying boundary condition $T(x,1) = x$ for any $x \in [0,1]$.

[2] This assumption is not generally accepted, e.g., in [4] it is omitted.

Definition 1 Let X_K and X_L ($K \cap L = \emptyset$) be two groups of variables with values in $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. Then the *conditional basic assignment* of X_K given $X_L \in B \subseteq \mathbf{X}_L$ (for B such that $m^{\downarrow L}(B) > 0$) is defined as follows:

$$m_{X_K|X_L}(A|B) = \frac{\sum\limits_{\substack{C \subseteq \mathbf{X}_{K\cup L}: \\ C^{\downarrow K}=A \& C^{\downarrow L}=B}} m(C)}{m^{\downarrow L}(B)}$$

for any $A \subseteq \mathbf{X}_K$.

2.4 Credal sets

Even more general is the theory of credal sets [7]. A *credal set* $\mathcal{M}(X_K)$ describing a group of variables X_K is defined as a closed convex set of probability measures describing the values of these variables.[3]

In order to simplify the expression of operations with credal sets, it is often considered [12] that a credal set is the set of probability distributions associated to the probability measures in it. Under such consideration a credal set can be expressed as a *convex hull* of its extreme distributions

$$\mathcal{M}(X_K) = \mathrm{CH}\{\mathrm{ext}(\mathcal{M}(X_K))\}.$$

Consider a credal set describing X_K, i.e. $\mathcal{M}(X_K)$. For each $L \subset K$ its *marginal credal set* $\mathcal{M}(X_L)$ is obtained by element-wise marginalization, i.e.

$$\mathcal{M}(X_L) = \mathrm{CH}\{P^{\downarrow L} : P \in \mathrm{ext}(\mathcal{M}(X_K))\},$$

where $P^{\downarrow L}$ denotes the marginal distribution of P on $\mathbf{X}_L$.

Again, having two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing X_K and X_L, respectively (assuming that $K, L \subseteq N$), we say that these credal sets are *projective* if

$$\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L}). \tag{1}$$

Let us note that if K and L are disjoint, then $\mathcal{M}_1$ and $\mathcal{M}_2$ are always projective, as $\mathcal{M}_1(X_\emptyset) = \mathcal{M}_2(X_\emptyset) \equiv 1$.

Conditional credal sets are obtained from the joint ones by point-wise conditioning of the extreme points and subsequent linear combination of the resulting conditional distributions. More formally: Let $\mathcal{M}(X_{K\cup L})$ ($K \cap L = \emptyset$) be a credal set describing (groups of) variables $X_{K\cup L}$. Then for any $x_L \in \mathbf{X}_L$

$$\mathcal{M}(X_K|x_L) = \mathrm{CH}\{P(X_K|x_L) : P \in \mathrm{ext}(\mathcal{M}(X_{K\cup L}))\},$$

is a *conditional credal set* describing X_K given $X_L = x_L$.

[3] For $K = \emptyset$ let us set $\mathcal{M}(X_\emptyset) \equiv 1$.

3 Compositional models

Now, we are ready to recall compositional models in the four above-characterized frameworks.

3.1 Composition operator

The most important concept of this contribution is that of the *composition* operator, in any of the above-mentioned frameworks. First we will recall its probabilistic form [9], as the remaining ones were designed to have analogous properties.

3.1.1 Composition of probability distributions

Let P_1 and P_2 be two probability distributions of (groups of) variables X_K and X_L; then

$$(P_1 \triangleright P_2)(X_{K\cup L}) = \frac{P_1(X_K)\cdot P_2(X_L)}{P_2(X_{K\cap L})}, \tag{2}$$

whenever $P_1(X_{K\cap L}) \ll P_2((X_{K\cap L})$; otherwise, it remains undefined.

3.1.2 Composition of possibility distributions

Composition operator for possibility distribution was introduced in [14]. Since conditioning in possibilistic framework is dependent on the selected t-norm, it is quite natural that also the composition operator is t-norm dependent.

Considering a continuous t-norm T, two subsets $K, L \subset N$ (this time not necessarily disjoint) and two normalized possibility distributions $\pi_1(x_K)$ and $\pi_2(x_L)$, we define the *composition operator* of possibilistic distributions as an analogy with (2) by the expression

$$(\pi_1 \triangleright_T \pi_2)(x_K, x_L) = T\left(\pi_1(x_{I_1}), \pi_2(x_{I_2}) \triangle_T \pi_2(x_{I_1\cap I_2})\right).$$

In contrast with the probabilistic case in possibility theory the composition operator is always defined.

3.1.3 Composition of basic assignments

Evidential compositional models are based on the concept of the operator of composition of basic assignments, introduced in [11] in the following way.

Definition 2 For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ a *composition* $m_1 \triangleright m_2$ is defined for all $C \subseteq \mathbf{X}_{K\cup L}$ by one of the following expressions:

(a) if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$ and $C = C^{\downarrow K} \bowtie C^{\downarrow L}$ then

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K})\cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})};$$

(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

$$(m_1 \triangleright m_2)(C) = 0.$$

Again, in this framework the composition operator is always defined.

3.1.4 Composition of projective credal sets

The compositional models for credal sets are under development — at present we only deal with the composition of projective credal sets [17].

Definition 3 For two projective credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing X_K and X_L, a *composition* $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is defined by the following expression:

$$\begin{aligned}(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) \quad = \quad & \mathrm{CH}\{(P_1 \cdot P_2)/P_2^{\downarrow K\cap L} : P_1 \in \mathrm{ext}(\mathcal{M}_1(X_K)),\\ & P_2 \in \mathrm{ext}(\mathcal{M}_2(X_L)), P_1^{\downarrow K\cap L} = P_2^{\downarrow K\cap L}\}.\end{aligned}$$

In all settings the resulting model keeps the first marginal, in case of projective distributions or basic assignments both [9, 14, 11]. As the credal composition operator is defined only for projective credal sets, it is not surprising, that it also keeps both marginals [17].

3.2 Perfect sequences and their properties

In this paragraph we will recall repetitive application of the composition operator with the goal to create a multidimensional model. As the theory of credal sets is the most general among theories studied in this contribution,[4] we will present all concepts and results in this section for credal sets.

Since the operator is not associative [9, 14, 11, 15], we have to specify in which order the low-dimensional credal sets are composed together. To make the formulae more transparent we will omit parentheses in case that the operator is to be applied from left to right, i.e., in what follows

$$\begin{aligned}&\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 \triangleright \cdots \triangleright \mathcal{M}_{m-1} \triangleright \mathcal{M}_m \\ &\qquad = (\cdots((\mathcal{M}_1 \triangleright \mathcal{M}_2) \triangleright \mathcal{M}_3) \triangleright \cdots \triangleright \mathcal{M}_{m-1}) \triangleright \mathcal{M}_m.\end{aligned} \tag{3}$$

Furthermore, we will always assume $\mathcal{M}_i$ be a credal set describing X_{K_i} and call $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \ldots, \mathcal{M}_m$ *generating sequence* of the model (3).

The reader familiar with some papers on compositional models knows that one of the most important notions of this theory is that of a so-called perfect sequence.

[4] The only exception is possibility theory, where the composition operator is parameterized by a continuous t-norm, and so is the concept of perfectness.

Definition 4 A generating sequence of credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ is called *perfect* if

$$\begin{aligned} \mathcal{M}_1 \triangleright \mathcal{M}_2 &= \mathcal{M}_2 \triangleright \mathcal{M}_1, \\ \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 &= \mathcal{M}_3 \triangleright (\mathcal{M}_1 \triangleright \mathcal{M}_2), \\ &\vdots \\ \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m &= \mathcal{M}_m \triangleright (\mathcal{M}_1 \triangleright \cdots \triangleright \mathcal{M}_{m-1}). \end{aligned}$$

It is evident that the necessary condition for perfectness is pairwise projectivity (i.e. (1) holds for any pair of credal sets from the generating sequence in question) of low-dimensional credal sets. However, it need not be sufficient.

Therefore a stronger, necessary and sufficient condition, expressed by the following assertion [15], must be fulfilled.

Lemma 1 *A generating sequence $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ is perfect* iff *the pairs of credal sets $\mathcal{M}_j$ and $(\mathcal{M}_1 \triangleright \cdots \triangleright \mathcal{M}_{j-1})$ are projective, i.e. if*

$$\mathcal{M}_j(X_{K_j \cap (K_1 \cup \cdots \cup K_{j-1})}) = (\mathcal{M}_1 \triangleright \cdots \triangleright \mathcal{M}_{j-1})(X_{K_j \cap (K_1 \cup \cdots \cup K_{j-1})}),$$

for all $j = 2, 3, \ldots, m$.

From Definition 4 one can hardly see what are the properties of the perfect sequences besides the algebraic ones; the most important one is expressed by the following characterization theorem [15], which also suggests why these sequences are called perfect.

Theorem 1 *A generating sequence of credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ is perfect* iff *all the credal sets from this sequence are marginal to the composed credal set $\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m$:*

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m)(X_{K_j}) = \mathcal{M}_j(X_{K_j}),$$

for all $j = 1, \ldots, m$.

The following theorem [17] is quite special within this contribution, as it deals with the relationship between perfect sequences in different frameworks. Nevertheless, it is not only an interesting result, but also an effective tool for the proof of a part of Theorem 3 in the next section.

Theorem 2 *Let $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ be a perfect sequence of credal sets such that each $\mathcal{M}_i, i = 1, \ldots m$, is the convex hull of its extreme points, i.e.,*

$$\mathcal{M}_i(X_{K_i}) = \mathrm{CH}\{P_i : P_i \in \mathrm{ext}(\mathcal{M}_i(X_{K_i}))\}.$$

Then

$$\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m$$

is a convex hull of all

$$P_1 \triangleright P_2 \triangleright \ldots \triangleright P_m$$

such that each $P_i \in \mathrm{ext}(\mathcal{M}_i(X_{K_i}))$, and $P_1, P_2, \ldots, P_m$ form a perfect sequence.

4 Networks

In this section we will deal with networks in each of the theories listed in Section 2 and study their relationship with the corresponding compositional models. First let us recall the basic concepts of particular networks.

4.1 Bayesian networks

As already mentioned in the introduction, Bayesian networks are probably the most popular representative of Graphical Markov models.

Relationships among variables in Bayesian networks are determined in two ways. Structural information describing the existence of a "direct" dependence of variables is given by a graph, while the quantitative information is given by a system of conditional probability distributions. Thus, a Bayesian network is a couple: an *acyclic directed graph* and a *system of conditional probability distributions*. In this system there are as many distributions as variables, i.e. nodes of the graph. For each variable there is a conditional distribution given all *parent*[5] variables in the condition. Some of nodes (at least one because of acyclicity) are parentless and their distributions are in fact unconditional.

Let us denote $\mathcal{BN}(X_N)$ the class of Bayesian networks over X_N.

4.2 Possibilistic networks

Possibilistic networks (Benferhat et al. [3] call them *directed possibilistic graphs*) can be introduced as a possibilistic counterpart of Bayesian networks in the following way:

Relationships among variables in possibilistic belief networks are determined analogous to Bayesian networks: an acyclic directed graph and a *system of conditional possibility distributions*. Nevertheless, there is one more parameter, a continuous t-norm (frequently minimum or product). So, let us denote $\Pi\mathcal{N}_T(X_N)$ the class of possibilistic networks over X_N (with respect to t-norm T).

4.3 Evidential networks

Evidential networks were introduced in [16] as a concept derived from perfect sequences, analogous to Bayesian networks: an acyclic directed graph and a *system of conditional basic assignments*. Conditional basic assignments are defined in accordance with Definition 1 only for focal elements in the condition. Again, let us denote $\mathcal{EN}(X_N)$ the class of possibilistic networks over X_N.

There exists also an alternative definition of networks in evidence theory, so-called *directed evidential networks* [5], but these models can hardly be considered to be a counterpart of Bayesian networks, as the graph has completely different interpretation.

[5]Node i is a parent of node j in a graph if there is a directed edge leading from i to j. The set of parents of j will be denoted $pa(j)$.

4.4 Credal Networks

A *credal network* [1] over X_N is a pair $(\mathcal{G}, \{\mathbf{P}^1, \ldots, \mathbf{P}^k\})$ such that for any $i = 1, \ldots, k$ $(\mathcal{G}, \mathbf{P}^i)$ is a Bayesian network over X_N, i.e. any $\mathbf{P}^i$ is a system of conditional probability distribution forming the joint distribution of X_N, $P^i(X_N)$.

The resulting model is a credal set, which is the convex hull of the Bayesian networks, i.e.

$$\mathrm{CH}\{P^1(X_N), \ldots P^k(X_N)\}.$$

It is evident, that this definition looses the attractiveness of Bayesian networks, where the overall information is computed from the local pieces of information. Let us denote by $\mathcal{CN}(X_N)$ the class of credal networks over X_N.

The most popular (and also most effective) type of credal networks are those called separately specified. A *separately specified credal network* over X_N is a pair $(\mathcal{G}, \mathbf{M})$, where $\mathbf{M}$ is a set of conditional credal sets $\mathcal{M}(X_i|pa(X_i))$ for each $X_i \in X_N$ and $pa(X_i)$ denotes the set of parent variables of X_i. Here the overall model is obtained analogous to Bayesian networks as the strong extension of the $\mathcal{M}(X_i|pa(X_i)), i \in N$. Analogous to previous paragraph let us denote by $\mathcal{SCN}(X_N)$ the class of separately specified credal networks over X_N.

Nevertheless, there exist a lot of situations in which separately specified credal networks cannot be used or their use leads to less specific models. For more details the reader is referred to [1].

4.5 Networks and Prefect Sequences

In this subsection we will overview, the relationship between networks and perfect sequences. For this purpose let us denote by $\mathcal{PPS}(X_N)$, $\Pi\mathcal{PS}_T(X_N)$, $\mathcal{EPS}(X_N)$ and $\mathcal{CPS}(X_N)$ the classes of perfect sequences[6] over X_N of probability distributions, possibility distributions, basic assignments and credal sets, respectively.

Theorem 3 *For any* X_N

(i) $\mathcal{BN}(X_N) = \mathcal{PPS}(X_N)$,

(ii) $\Pi\mathcal{N}_T(X_N) = \Pi\mathcal{PS}_T(X_N)$,

(iii) $\mathcal{EN}(X_N) \subset \mathcal{EPS}(X_N)$,

(iv) $\mathcal{SCN}(X_N) \subset \mathcal{CPS}(X_N) \subset \mathcal{CN}(X_N)$.

The proofs of particular parts of theorem can be found in [10, 16, 17] and for the description of an algorithm reconstructing a network from a perfect sequence the reader is referred to the same papers.

The fact, that compositional models for credal sets are based on "local knowledge" even in cases, when the credal network is not separately specified, can be considered as an advantage of these models.

[6] In case of possibility distributions T-perfect sequences.

5 Conclusions

The aim of this contribution was to overview relationships between compositional models and networks in four specific settings of imprecise probabilities.

In probability theory, special class of these models, called perfect sequences, was proved to be equivalent to Bayesian networks in such a sense, that any Bayesian network can be expressed as a composition of a perfect sequence of probability distributions and vice versa. In possibility theory an analogous equivalence relation holds true: for identical t-norms and a suitable choice of a conditioning rule.

On the contrary in evidence theory the equivalence is no more valid. Any evidential network can be expressed in the form of a compositional model, but not vice versa. In other words, the class of compositional models is much richer. In credal sets theory, the relationship is even more complicated: any separately specified credal network can be expressed in the form of a compositional model and any compositional model can be expressed in the form of a credal network. The reverse implications do not hold.

From the results presented in this contribution it is evident, that both evidential and credal compositional models are worth-developing, as they are more flexible than evidential networks and separately specified credal networks, respectively.

Acknowledgment

This work was supported by the Czech Science Foundation through Project 13-20012S.

References

[1] A. Antonucci and M. Zaffalon, Decision-theoretic specification of credal networks: a unified language for uncertain modeling with sets of Bayesian networks. *Int. J. Approx. Reasoning,* **49** (2008), pp. 345–361.

[2] C. Beeri, R. Fagin, D. Maier, M. Yannakakis, On the desirability of acyclic database schemes, *J. of the Association for Computing Machinery,* **30** (1983), pp. 479–513.

[3] S. Benferhat, D. Dubois, L. Gracia and H. Prade, Directed possibilistic graphs and possibilistic logic. In: B. Bouchon-Meunier, R.R. Yager, (eds.) *Proc. IPMU'98*, Editions E.D.K. Paris, pp. 1470–1477.

[4] B. Ben Yaghlane, Ph. Smets, K. Mellouli, Belief functions independence: I. the marginal case. *Int. J. Approx. Reasoning*, **29** (2002), 47–70.

[5] B. Ben Yaghlane, Ph. Smets and K. Mellouli, Directed evidential networks with conditional belief functions. *Proceedings of ECSQARU 2003,* eds. T. D. Nielsen, N. L. Zhang, 291–305.

[6] G. de Cooman, Possibility theory II: Conditional possibility. *Int. J. of General Systems* **25** (1997), pp. 325–351.

[7] F. G. Cozman, Sets of probability distributions, independence, and convexity, *Synthese*, **186** (2012), pp. 577–600.

[8] M. Daniel, Belief conditioning rules for classic belief functions, *Proceedings of WUPES'09,* eds. T. Kroupa, J. Vejnarová, pp. 46–56.

[9] R. Jiroušek, Composition of probability measures on finite spaces. *Proc. of UAI'97*, (D. Geiger and P. P. Shenoy, eds.). Morgan Kaufmann Publ., San Francisco, California, pp. 274–281.

[10] R. Jiroušek and J. Vejnarová, Construction of multidimensional models by operators of composition: current state of art. *Soft Computing,* **7** (2003), pp. 328–335.

[11] R. Jiroušek, J. Vejnarová, M. Daniel, Compositional models for belief functions. In: De Cooman G, Vejnarová J, Zaffalon M (eds.) *Proceedings of ISIPTA'07.* Praha, pp. 243–252.

[12] S. Moral and A. Cano, Strong conditional independence for credal sets, *Ann. of Math. and Artif. Intell.*, **35** (2002), pp. 295–321.

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

[14] J. Vejnarová, Composition of possibility measures on finite spaces: preliminary results. In *Proc. IPMU'98*, Paris, France, (B. Bouchon-Meunier, R. R. Yager, eds.). E.D.K., Paris, 1998, pp. 25–30.

[15] J. Vejnarová, Operator of composition for credal sets, *ISIPTA'13: 8th International Symposium on Imprecise Probability: Theories and Applications*, pp. 355–364.

[16] J. Vejnarová, A Comparison of Evidential Networks and Compositional Models, *Kybernetika* **50** (2014), pp. 246–267.

[17] J. Vejnarová, Credal Compositional Models and Credal Networks. In: T. Augustin, S. Doria, E. Miranda, E Quaeghebeur (eds.) *Proceedings of ISIPTA'15.* Rome, pp. 315–324.

Influence diagrams for speed profile optimization: computational issues

Jiří Vomlel and Václav Kratochvíl
Institute of Information Theory and Automation
Czech Academy of Sciences
vomlel@utia.cas.cz, velorex@utia.cas.cz

Abstract

Influence diagrams were applied to diverse decision problems. However, the general theory is still not sufficiently developed if the variables are continuous or hybrid and the utility functions are nonlinear. In this paper, we study computational problems related to the application of influence diagrams to vehicle speed profile optimization and suggest an approximation of the nonlinear utility functions by piecewise linear functions.

1 Introduction

In this paper, we use an example inspired by a real problem – a car moving on a road – to study various issues related to computations with influence diagrams. The modeled car is equipped with an automatic transmission and its speed is controlled using the throttle and the brakes. There are various speed limits on the road (e.g., 130 km/h on a highway or 50 km/h in an urban area). The goal is to find an optimal strategy for passing the road while minimizing (*i*) time spend on the road, (*ii*) the fuel consumption, or (*iii*) a mixture of both.

There are two principal ways for representing the solution:

speed profile – a function that assigns a speed value to all points on the road,

control policy – a function that assigns control values of the throttle and the brakes for every possible speed and to every point on the road.

The *control policy* is more general. In case of the *speed profile* the vehicle uses an additional regulator that follows the speed profile by controlling the car acceleration using the throttle and the brakes. In the *control policy*, the control signals are already precomputed for all admissible speed values. This becomes especially handy in real situations when the driver has to suddenly slow down or even to stop due to an unexpected traffic situation and the precomputed speed profile becomes obsolete.

Since all variables (speed, acceleration, throttle, brakes) are continuous by their nature, it would be natural to work with continuous or hybrid influence diagrams. Unfortunately, the theory of continuous influence diagrams is not sufficiently developed (especially for nonlinear utility functions). In this paper we perform experiments with discrete influence diagrams. One of our goals is to analyze the shape of nonlinear relations and propose good approximations.

An influence diagram (Howard and Matheson, 1981) is a Bayesian network augmented with decision variables and utility functions. Influence diagrams were applied to diverse decision problems. Recently, we introduced influence diagrams to the problem of optimization of a vehicle speed profile. We performed computational experiments in which an influence diagram was used to optimize the speed profile of a Formula 1 race car at the Silverstone F1 circuit (Kratochvíl and Vomlel, 2015).

In this paper we split the vehicle path into n segments of the same length s. For each segment of the vehicle path there are two random variables V_i and V_{i+1}, one decision variable U_i, and one utility potential f_{i+1}. In Figure 1, we present the structure of a part of the ID corresponding to one segment of the path. The values of i are from the set $\{1, 2, \ldots, n-1\}$. The physical model of the vehicle is given in

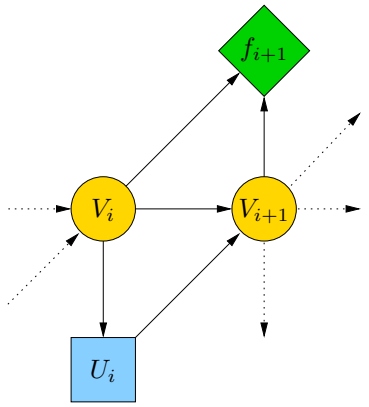


Figure 1: A part of the influence diagram for one path segment

Section 2. It is used to define the probability and utility functions of the influence diagram.

In this paper, we generally allow variables to be discrete or continuous and the main theoretical results presented in the paper are valid for both types of variables. However, experiments were performed with discrete variables only. For the sake of brevity we do not discuss related work in this paper – we refer interested readers to Kratochvíl and Vomlel (2015).

2 Vehicle physics

We model the vehicle behavior using the laws of physics. To model the engine behavior and the fuel consumption we assume the vehicle to be a passenger car and we follow the approach of Chang and Morlok (2005). The values of variables[1] describing the car state are defined by the following formulas[2].

Velocity at the coordinate $i+1$

$$v_{i+1} \quad = \quad v(a_i, v_i) \quad = \quad \sqrt{(v_i)^2 + 2 \cdot s \cdot a_i} \ , \tag{1}$$

where a_i and v_i is acceleration and velocity at the coordinate i, respectively. Let a_t^{max} be the maximum tangential acceleration of the vehicle,[3] a_t^{min} be the maximum tangential deceleration,[4] Engine acceleration at segment i is defined by the following equation:

$$a_i^e \quad = \quad a^e(u_i, v_i) \quad = \quad \begin{cases} u_i \cdot (a_t^{max} - c_a \cdot v_i) & \text{if } u_i > 0 \\ u_i \cdot a_t^{min} & \text{otherwise.} \end{cases} \tag{2}$$

In this paper we will consider a vehicle with values $a_t^{max} = 4$, $c_a = 0.06$, and $a_t^{min} = 5$. u_i is the control at the coordinate i. It has values from $\langle 0, 1 \rangle$ where negative ones correspond to braking, positive ones to using throttle. Deceleration caused by friction forces and aerodynamic drag is

$$a_i^d \quad = \quad a^d(v_i) \quad = \quad c_r + c_v \cdot (v_i)^2 \tag{3}$$

where $c_r = 0.1273$ and $c_v = 0.000257$ for the considered vehicle. Acceleration at segment $[i, i+1]$ is

$$a_i \quad = \quad a(u_i, v_i) \quad = \quad a^e(u_i, v_i) - a^d(v_i) \ . \tag{4}$$

By putting equations (1)–(4) altogether we get

$$\begin{aligned} v_{i+1} \quad &= \quad v'(u_i, v_i) \\ &= \quad \begin{cases} \sqrt{(v_i)^2 + 2 \cdot s \cdot \left(u_i \cdot (a_t^{max} - c_a \cdot v_i) - c_r - c_v \cdot (v_i)^2 \right)} & \text{if } u_i > 0 \\ \sqrt{(v_i)^2 + 2 \cdot s \cdot \left(u_i \cdot a_t^{min} - c_r - c_v \cdot (v_i)^2 \right)} & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

Time spent at the path segment $[i, i+1]$

$$t_{i+1} \quad = \quad t(v_i, v_{i+1}) \quad = \quad s \cdot \left(\frac{v_i + v_{i+1}}{2} \right)^{-1} \ . \tag{6}$$

[1] We use the symbol without subscript to denote the function that specifies the variable's value.

[2] Note that the relations between variables follow the edges of the influence diagram from Figure 1

[3] It is a property of the vehicle engine (without considering the aerodynamic drag and friction forces). The real maximum acceleration is lower.

[4] It is a property of the vehicle brakes (without considering the aerodynamic drag and friction forces). The real maximum deceleration is higher.

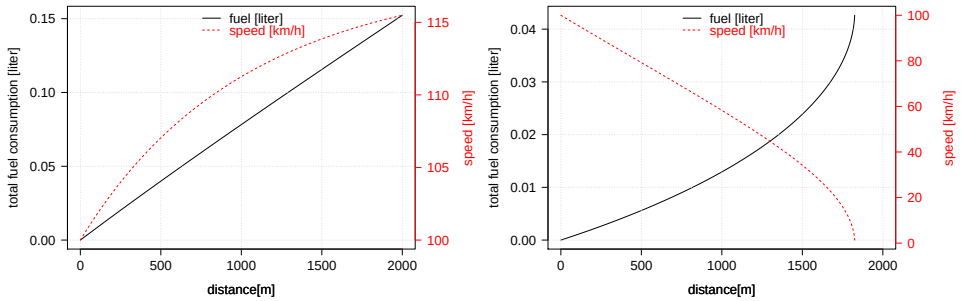


Figure 2: The total fuel consumption and the speed with the initial speed 100 km/h and the control $u = 0.2$ and $u = 0$, respectively.

When modeling the fuel consumption, we assume it is proportional to the work done by the engine (Chang and Morlok, 2005), which is the acceleration multiplied by the vehicle mass and by the distance s, plus a low fuel consumption constant per time:

$$\begin{aligned} g_{i+1} &= g(v_i, v_{i+1}) \\ &= c_g \cdot s \cdot m \cdot \max\left\{0, \frac{(v_{i+1})^2 - (v_i)^2}{2s} + a_d(v_i)\right\} + g^{min} \cdot t(v_i, v_{i+1}) \ , \end{aligned} \quad (7)$$

where the considered constants are the vehicle mass in kilograms $m = 1759$, the fuel rate in liter per one Joul of energy $c_g = 10^{-7}$, and the constant fuel consumption in liter per second $g^{min} = 1/3600$. The vehicle behavior in terms of the fuel consumption and its speed is illustrated in Figure 2.

3 Speed constraints in the model

We assume that a maximum speed v_i^{max} and a minimum speed v_i^{min} is given in advance at each path coordinate $i = 1, \ldots, n$. Let $\mathcal{V}_i$ denote the set of admissible speed values at i and let the admissible set at the end of the path be

$$\mathcal{V}_n = \{v \in \mathcal{V}, v_n^{min} \leq v \leq v_n^{max}\} \ . \quad (8)$$

We apply the constraints during optimization process where we allow to select only those control signals $u_i \in \mathcal{U}$ that lead to $v_{i+1} = v'(u_i, v_i)$ belonging to $\mathcal{V}_{i+1}$. We define functions $\mathcal{U}_i(V_i)$ that for each value v_i of variable V_i provide the set of admissible control values:

$$\mathcal{U}_i(v_i) = \{u \in \mathcal{U} : v'(u_i, v_i) \in \mathcal{V}_{i+1}\} \ . \quad (9)$$

This set inductively defines the set of admissible speed values at i for which there exist an admissible control value:

$$\mathcal{V}_i = \{v \in \mathcal{V} : v_i^{min} \leq v \leq v_i^{max}, \ \mathcal{U}_i(v) \neq \emptyset\} \ . \quad (10)$$

This, again, inductively defines set $\mathcal{U}_{i-1}(v_{i-1})$. This process is repeated until $i = 1$.

4 Expected utility of a control policy

In the sequel we will use the following abbreviations

$$\begin{aligned} \sum_{V_i} \varphi(V_i, \cdot) &= \sum_{v_i \in \mathcal{V}_i} \varphi(V_i = v_i, \cdot) \text{ and} \\ \max_{U_i} \psi(U_i, V_i) &= \max_{u_i \in \mathcal{U}_i(v_i)} \psi(U_i = u_i, V_i = v_i) \ . \end{aligned}$$

$\mathcal{M}$ will be a generalized marginalization operation. The operator $\mathcal{M}$ acts differently for a discrete random variable A, a continuous random variable B, and a decision variable U of a (probability or utility) potential ψ:

$$\begin{aligned} \underset{A}{\mathcal{M}}\ \psi(A, \ldots) &= \sum_{A} \psi(A, \ldots), \qquad \underset{B}{\mathcal{M}}\ \psi(B, \ldots) = \int \psi(B = b, \ldots)\ db, \\ \underset{U}{\mathcal{M}}\ \psi(U, \ldots) &= \max_{U}\ \psi(U, \ldots) \ . \end{aligned}$$

The control of the vehicle speed will be realized by means of the control policy.

Definition 1. Control policy is a set of functions

$$\delta = \left\{\delta(U_i|V_i) : \ i \in \{1, \ldots, n-1\}, v_i \in \mathcal{V}\right\}$$

such that for all $i = 1, \ldots, n$ and all $v_i \in \mathcal{V}$ it maps $u_i \in \mathcal{U}$ to values from $[0, 1]$ and it holds that

$$\sum_{u_i \in \mathcal{U}} \delta(U_i = u_i | V_i = v_i) = 1 \ . \tag{11}$$

Definition 2. A control policy δ is deterministic if for all $i = 1, \ldots, n$ and all $v_i \in \mathcal{V}$ it holds that there is a function $u_i : \mathcal{V} \to \mathcal{U}$ such that for all $u \in \mathcal{U}$

$$\delta(U_i = u | V_i = v_i) = \begin{cases} 1 & \text{if } u = u_i(v_i) \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Remark 1. In this paper, all considered policies will be deterministic.

Definition 3. The expected value E_f of a deterministic control policy δ specified by functions u_i is the sum or the integral over all possible configurations of random variables of the products of the probability and the criteria value of that configuration:

$$E_f(\delta) = \underset{V_1, \ldots, V_n}{\mathcal{M}} P(V_1, \ldots, V_n) \cdot f(V_1, \ldots, V_n) \tag{13}$$

where

$$P(V_1, \ldots, V_n) = P(V_1) \cdot \prod_{i=1}^{n-1} P(V_{i+1} | U_i = u_i(v_i), V_i) \tag{14}$$

$$f(V_1, \ldots, V_n) = \sum_{i=1}^{n-1} f(V_i, V_{i+1}) \ . \tag{15}$$

The criteria to be optimized will be the expected value E_f of a deterministic control policy.

Definition 4. An optimal deterministic policy δ^* is a deterministic policy such that it holds for all control policies δ that

$$E_f(\delta) \quad \leq \quad E_f(\delta^*) \ . \tag{16}$$

We will use symbol u_i^* to denote the function $u_i : \mathcal{V} \rightarrow \mathcal{U}$ that specifies the optimal deterministic policy δ^* according to Definition 2. The symbol $u_i^*(V_i)$ denotes the set of functions u_i^* for all values v_i of variable V_i.

Using the recursive application of the commutative and distributive laws we get the following theorem that specifies a computationally efficient algorithm for finding an optimal decision policy. Note that our algorithm is just a special case of general inference methods for influence diagrams (Jensen et al., 1994; Shenoy, 1992; Shachter and Peot, 1992). But since our influence diagram has a simple structure it is useful to derive a simple inference algorithm tailored for the task we solve. Note that, in this case, the algorithm does not involve divisions. The computations can be also viewed as a special case of dynamic programming (Bellman, 1957).

Theorem 1.

$$E_f^* \quad = \quad E_f(\delta^*) \quad = \quad \underset{V_1}{\mathcal{M}} \ P(V_1) \cdot \psi(V_1) \ , \tag{17}$$

where $\psi(V_1)$ *is computed recursively for* $i = 1, \ldots, n-1$ *as*

$$\psi(V_i) \quad = \quad \max_{U_i} \underset{V_{i+1}}{\mathcal{M}} \ P(V_{i+1}|V_i, U_i) \cdot \Big(f(V_i, V_{i+1}) + \psi(V_{i+1}) \Big) \ . \tag{18}$$

with the recursion terminal values being $\psi(V_n) = \mathbb{0}(V_n)$*, where* $\mathbb{0}(V_n)$ *stands for the vector taking for all states of variable* V_n *value zero.*

The proof can be found in Appendix A.

Remark 2. In each step $i = 1, \ldots, n$, an optimal deterministic policy is specified (according to Definition 2) by a function $u_i : \mathcal{V} \rightarrow \mathcal{U}$ such that $u_i(v_i) = u_i^*(v_i)$, where $u_i^*(v_i)$ is a value of U_i that maximize formula (18) for a given v_i .

5 Deterministic continuous model for the total time minimization

In this section we will present a special case for which it is easy to find an optimal speed profile even if all variables are continuous. The optimality criteria will be the total time $\sum_{i=1}^{n-1} t(v_i, v_{i+1})$ and the goal will be to minimize it.

Definition 5. Let $v'(u_i, v_i)$ be the function specified in (5). If for $i = 1, \ldots, n-1$ it holds that

$$P(V_{i+1} = v_{i+1} | U_i = u_i, V_i = v_i) \quad = \quad \begin{cases} 1 & \text{if } v_{i+1} = v'(u_i, v_i) \\ 0 & \text{otherwise.} \end{cases}$$

then we say that the vehicle behavior is deterministic.

Next we present a corollary of Theorem 1 that specifies an algorithm for the case of a deterministic vehicle behavior.

Corollary 1. *Assume that the vehicle behavior is deterministic. Then*

$$E_f^* \quad = \quad E_f(\delta^*) \quad = \quad \underset{V_1}{\mathcal{M}} \; P(V_1) \cdot \psi(V_1) \;\; , \tag{19}$$

where $\psi(V_1)$ is computed recursively for $i = 1, \ldots, n-1$ and for all $v_i \in \mathcal{V}$ as:

$$\psi(v_i) \quad = \quad f\left(v_i, v'\big(\max \; \mathcal{U}_i(v_i), v_i\big)\right) + \psi\left(v'\big(\max \; \mathcal{U}_i(v_i), v_i\big)\right) \;\; . \tag{20}$$

The recursion terminal values are defined as $\psi(v_n) = 0$ for all $v_n \in \mathcal{V}$.

Proof. Formula (20) follows from (18) - the considered criteria is the minimization of the total time. Therefore $\max_{U_i}$ corresponds to picking the highest value from $\mathcal{U}_i(v_i)$. Also, note that for the deterministic vehicle behavior and for any potential $\xi(V_i, V_{i+1})$ it holds for all $u_i \in \mathcal{U}, v_i \in \mathcal{V}$ that

$$\underset{V_{i+1}}{\mathcal{M}} \; P(V_{i+1} | U_i = u_i, V_i = v_i) \cdot \xi(V_i = v_i, V_{i+1}) \quad = \quad \xi(V_i = v_i, V_{i+1} = v'(u_i, v_i)) \;\; .$$

□

From Corollary 1 we derive computationally efficient Algorithm 1 that can be used to compute efficiently the optimal speed profile of the vehicle satisfying the speed constraints. We will use function $w(u_i, v_{i+1})$ that gives the initial speed v_i such that after driving distance s with the control u_i the speed is v_{i+1}. The idea behind the algorithm is that the function f, which is is to be maximized, implies that the best policy for any $v_i, i = 1, \ldots, n-1$ is to speedup as much as possible to be able to slow down by maximum allowed decceleration to satisfy that $v_j^* \leq v_j^{max}$ for all $j > i$.

First, the maximal speed profile is constructed from the speed constraints and the maximum decceleration of the vehicle. Second, the best policy is found with the maximum acceleration until the speed meets the maximum profile constructed in the first stage of the algorithm.

6 Experiments

In the experiments, we considered the speed and control variables to be discrete, i.e. sets $\mathcal{V}$,$\mathcal{U}$ are finite with the discretization steps being d_V, d_U, respectively. In

input : $v_i^{max}, i = 1, \ldots, n$ – maximal speed values
output: $v_i^*, i = 1, \ldots, n$ – speed values maximizing E_f (see Definition 3)

$v_n^* = v_n^{max}$;
for $i = n-1, \ldots, 1$ **do**
  $v_i^* = w(-1, v_{i+1}^*)$;
  if $(v_i^* > v_i^{max})$ **then**
    $v_i^* = v_i^{max}$;
  end
end
for $i = 1, \ldots, n-1$ **do**
  $v_{i+1} = v'(+1, v_i^*)$;
  if $(v_{i+1} < v_{i+1}^*)$ **then**
    $v_{i+1}^* = v_{i+1}$;
  end
end

Algorithm 1: Optimal speed profile construction for the deterministic vehicle behavior.

this case we use linear approximations of utility values of $v_i = v(u_{i-1}, v_{i-1})$, $v_i \notin \mathcal{V}$ by a mixture of utility values $\underline{v}_i \leq v_i$ and $\overline{v}_i \geq v_i$ that are the closest values from $\mathcal{V}$ to v_i. The mixture weights are the probabilities that are defined as

$$P(V_i = v | U_{i-1} = u_{i-1}, V_{i-1} = v_{i-1}) \quad = \quad \begin{cases} 1 - \dfrac{|v - v_i|}{d_V} & \text{for } v = \underline{v}_i, \overline{v}_i \\ 0 & \text{otherwise.} \end{cases} \qquad (21)$$

To get an into the problem we performed the following computational experiment. Assume a road section of length 2 km in a flat area and the speed limit of 90 km/h in the whole section and with three short subpaths with the speed limit of 50 km/h. Let $s = 20$ m. The speed limit profile of the road can be seen in the upper part of Figure 3. The area of forbidden speeds is highlighted. The black line illustrates a speed profile of a car starting with initial speed of 80 km/h, following the control policy calculated using Theorem 1. The probability potentials were defined as in (21) and $\mathcal{V}$,$\mathcal{U}$ had 100 values.

Using deterministic relation between variables, we are inevitably working with states of zero probability. If the task is minimization of a criteria the zero probability values may lead to wrong solutions. Therefore we formulate the problem as a maximization task. Instead of the minimization of a specific mixture of the fuel consumption and the total time, we maximize the *savings* with respect to the worst performance. As the optimality criteria we use a mixture of the normalized total time savings and the normalized fuel savings. The normalized utility functions for the time and fuel savings at segment $[i, i+1]$ are defined as

$$f_{i+1}^t = f^t(v_i, v_{i+1}) = 1 - \frac{t(v_i, v_{i+1})}{t^{max}}, \qquad f_{i+1}^f = f^f(v_i, v_{i+1}) = 1 - \frac{g(v_i, v_{i+1})}{g^{max}},$$

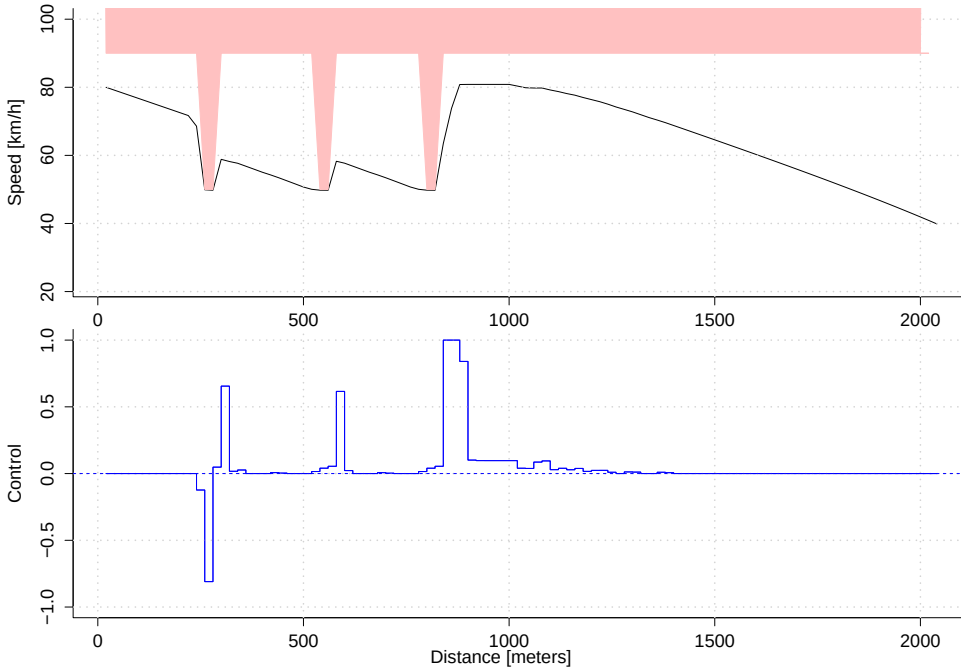


Figure 3: Generated speed profile and corresponding control profile

where $t(v_i, v_{i+1})$ and $g(v_i, v_{i+1})$ are defined by formula (6) and (7), respectively. t^{max} and g^{max} are the maximum possible time spent and fuel consumption in one segment. In the experiments we used utility function f defined (for $\alpha = 0.5$)

$$f \quad = \quad \sum_{i=1}^{n-1} \alpha f_i^t + (1-\alpha) f_i^f \quad . \tag{22}$$

Remark 3. For the speeds close to zero the values of $t(v_i, v_{i+1})$ and $g(v_i, v_{i+1})$ are very high. This would imply high values of t^{max} and g^{max}. Consequently, for most of other speed values the functions f_{i+1}^t and f_{i+1}^f would provide values close to one. This may cause rounding errors. To avoid this problem we disregard speeds lower than 4 km/h for the definitions of t^{max} and g^{max}.

In Figure 3, we present results of our numerical experiment. In the upper part the computed optimal speed profile is presented. The corresponding values of the control variable (the throttle or the brakes) are depicted in the lower part of the figure. It is interesting to note that most of the time the car is in a so called *flying mode*, which is driving with the neutral gear with no throttle or brakes. In case of a longer road without speed limits, the optimal speed stabilizes (for this settings) around 80 km/h - see the road section around 900 m. Because there is

no requirement on the speed at the end - the algorithm decided to enter the flying mode, similarly, as in the case when the vehicle is approaching 50 km/h speed limits.

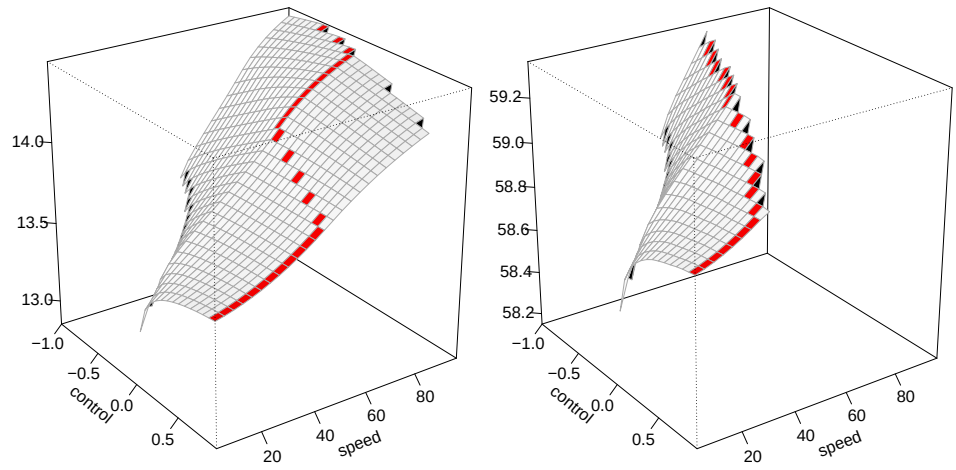
According to Theorem 1, the calculations are performed in the direction from the road's furthermost point backwards. The development of the values of the expected utility function in two randomly selected points of the path can be seen in Figure 4. Axis x and y correspond to speed and control, axis z refers to the expected utility. Figure 4a corresponds to the iteration 15 of the algorithm ($i = 87$), while Figure 4b corresponds to the iteration 62 ($i = 39$). Forbidden combinations of speed and control are not depicted. In Figure 4 the highlighted facets corresponds to maximal expected utility for the given speed. For every value of speed, we store respective control as the optimal deterministic policy in this points - see Remark 2.

From Figure 4a we can deduce that the best strategy for a low speed is to use the full throttle (first, to speed up and than to use the flying mode). For speeds of about $50 - 60$ km/h it starts to be better to use the flying mode immediately. For very high speeds, the optimal strategy has to be to use the brakes in order to satisfy the speed limits. The overall view of the image suggests that global optimum is at the highest speeds. It is logical, because with a high initial speed a lot of the fuel and time can be saved. Figure 4b corresponds to a driving situation just before reaching one of the speed limits of 50 km/h. Therefore more combinations of speed and control values are forbidden. However, the shape of the expected utility function is similar.

Remark 4. Note the scale of axis z in Figures 4a and 4b. Recall that, in every point, we are using a weighted mixture of normalized utility functions with values from interval $\langle 0, 1\rangle$. By maximization, we usually select combinations with values close to 1 and that is why the values of expected utility corresponds well to the number of the current algorithm iteration.

Our future goal is to move from discrete variables to the continuous ones. Therefore, it is interesting to see the shape of the expected utility function with respect to the control value and for a given speed. Let us take the utility function from Figure 4a and select five speed values. Respective slices are depicted in Figure 5. The gray solid lines show values from Figure 4a, the black lines show piecewise-linear approximations of each line. All approximations are composed from three lines. To find the best approximation of each curve, we used R package *segmented* (Muggeo, 2008). The package estimates linear and generalized linear models with one or more segmented relationships in the linear predictor. Estimates of the slopes and of the (possibly multiple) breakpoints are provided. In our experiments, we decided to fix the number of breakpoints to two and let the algorithm find their best positions.

In an influence diagram with continuous variables we would need to represent the optimal control policy at each step i by a function $u_i : \mathcal{V} \to \mathcal{U}$ (see Definition 2). The optimal control policy at point $i = 87$ of the path is depicted in Figure 6. We can see that piecewise linear functions may again represent good approximations.



(a) $i = 87$, ($1740m$, 15th iteration) (b) $i = 39$, ($780m$, 62nd iteration)

Figure 4: Expected utility as a function of values of variables V_i, U_i.

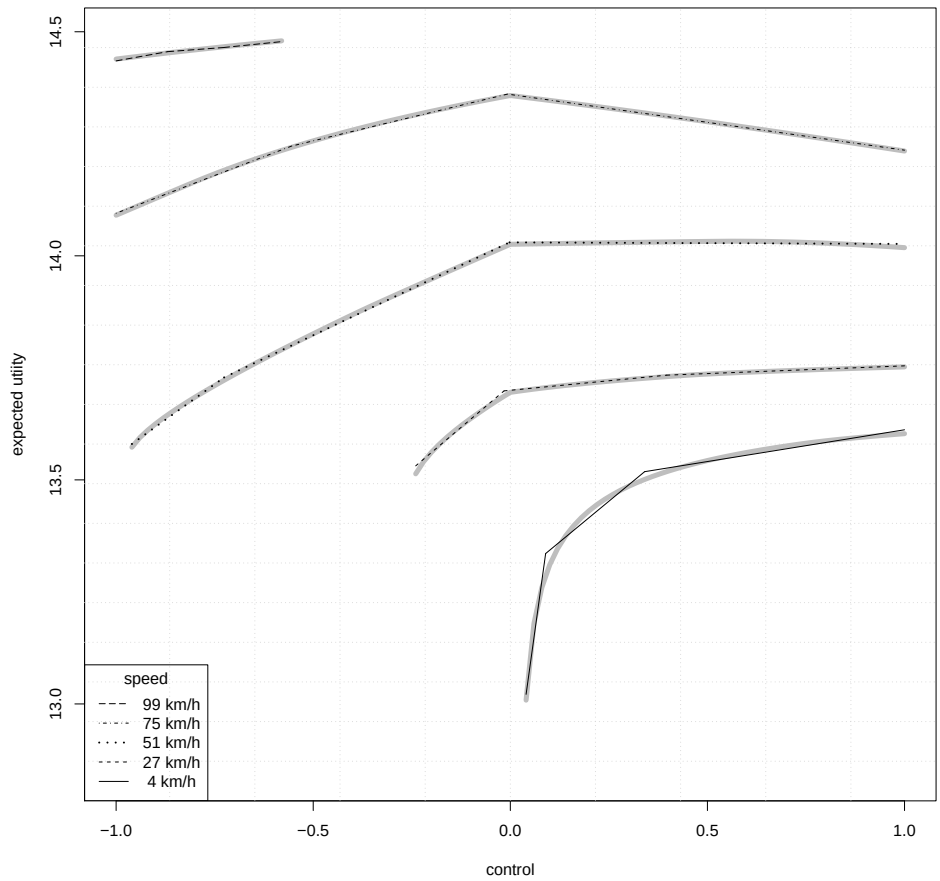


Figure 5: Expected utility as a function of the control values for several initial speed values, $i = 87$

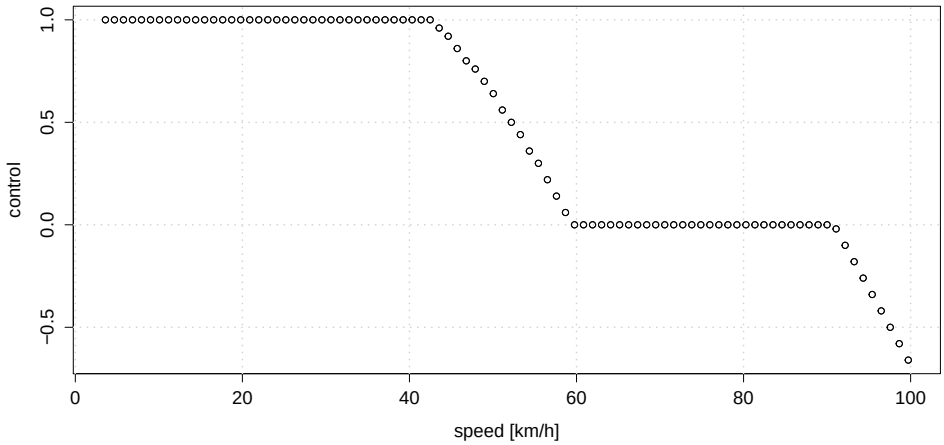


Figure 6: Optimal control policy, $i = 87$

7 Conclusions

We applied influence diagrams to optimization of a vehicle speed profile and performed numerical experiments on a 2-km-long path with few speed constraints. We considered optimality criteria based on a mixture of the fuel consumption and the total driving time. We derived the general inference algorithm for this type of influence diagrams and presented efficient modifications of this algorithm for specific cases. Finally, we used the numerical experiments to elicit the shape of expected utility and policy functions. In both cases piecewise linear functions seem to be good approximations that can be used in influence diagrams with continuous variables.

Acknowledgment

This work was supported by the Czech Science Foundation through Project 13-20012S.

References

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Chang, D. J. and Morlok, E. K. (2005). Vehicle speed profiles to minimize work and fuel consumption. *Journal of Transportation Engineering*, 131(3):173–182.

Howard, R. A. and Matheson, J. E. (1981). Influence diagrams. In Howard, R. A. and Matheson, J. E., editors, *Readings on The Principles and Applications of Decision Analysis*, volume II, pages 721–762. Strategic Decisions Group.

Jensen, F., Jensen, F. V., and Dittmer, S. L. (1994). From influence diagrams to junction trees. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 367–373. Morgan Kaufmann.

Kratochvíl, V. and Vomlel, J. (2015). Influence diagrams for the optimization of a vehicle speed profile. In *Proceedings of the Twelfth Annual Bayesian Modeling Applications Workshop*. Accepted for publication.

Muggeo, V. M. R. (2008). segmented: an R package to fit regression models with broken-line relationships. *R News*, 8(1):20–25.

Shachter, R. D. and Peot, M. A. (1992). Decision making using probabilistic inference methods. In *Proceedings of the Eighth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-92)*, pages 276–283, San Mateo, CA. Morgan Kaufmann.

Shenoy, P. P. (1992). Valuation based systems for Bayesian decision analysis. *Operations Research*, 40:463–484.

A Proof of Theorem 1

Proof. For any $j = 1, \ldots, n$ we will denote the joint probability distribution as

$$P(U_1, \ldots, U_j, V_1, \ldots, V_j) \quad = \quad P(V_1) \cdot \prod_{i=2}^{j} P(V_i|U_{i-1}, V_{i-1}) \cdot \delta(U_{i-1}|V_{i-1})$$

and the total utility as

$$f(V_1, \ldots, V_j) \quad = \quad \sum_{i=1}^{j-1} f(V_i, V_{i+1}) \ .$$

For the maximal expected value it holds that

$$\begin{aligned}
&E_f^* \\
&= \max_{U_1,\ldots,U_{n-1}} \mathop{\mathcal{M}}_{V_1,\ldots,V_n} \Big(P(U_1, \ldots, U_{n-1}, V_1, \ldots, V_n) \cdot f(V_1, \ldots, V_n) \Big) \\
&= \max_{U_1,\ldots,U_{n-1}} \mathop{\mathcal{M}}_{V_1,\ldots,V_n} \left(\begin{array}{l} P(U_1, \ldots, U_{n-1}, V_1, \ldots, V_n) \\ \cdot \Big(f(V_1, \ldots, V_n) + \psi(V_n) \Big) \end{array} \right) \qquad (23) \\
&= \max_{U_1,\ldots,U_{n-1}} \mathop{\mathcal{M}}_{V_1,\ldots,V_{n-1}} \left(\begin{array}{l} P(U_1, \ldots, U_{n-1}, V_1, \ldots, V_{n-1}) \\ \\ \cdot \displaystyle\sum_{V_n} P(V_n|V_{n-1}, U_{n-1}) \cdot \left(\begin{array}{l} f(V_1, \ldots, V_{n-1}) \\ +f(V_{n-1}, V_n) \\ +\psi(V_n) \end{array} \right) \end{array} \right) .
\end{aligned}$$

We can write

$$E_f^* = \max_{U_1,\ldots,U_{n-1}} \underset{V_1,\ldots,V_{n-1}}{\mathcal{M}} \left(\begin{array}{l} P(U_1,\ldots,U_{n-1},V_1,\ldots,V_{n-1}) \\ \cdot\Big(\xi(V_1,\ldots,V_{n-1}) + \psi(U_{n-1},V_{n-1})\Big) \end{array} \right) ,$$

where

$$\xi(V_1,\ldots,V_{n-1}) = \underset{V_n}{\mathcal{M}}\Big(P(V_n|V_{n-1},U_{n-1}) \cdot f(V_1,\ldots,V_{n-1})\Big) \quad (24)$$

$$\psi(U_{n-1},V_{n-1}) = \underset{V_n}{\mathcal{M}} P(V_n|V_{n-1},U_{n-1}) \cdot \Big(f(V_{n-1},V_n) + \psi(V_n)\Big) . \quad (25)$$

Equation (24) can be simplified to

$$\xi(V_1,\ldots,V_{n-1}) = \Big(\underset{V_n}{\mathcal{M}}\ P(V_n|V_{n-1},U_{n-1})\Big) \cdot f(V_1,\ldots,V_{n-1}) \quad (26)$$

$$= f(V_1,\ldots,V_{n-1}) , \quad (27)$$

where the second transformation is due to $\underset{V_n}{\mathcal{M}}\ P(V_n|V_{n-1},U_{n-1}) = 1$. This implies

$$E_f^* = \max_{U_1,\ldots,U_{n-1}} \underset{V_1,\ldots,V_{n-1}}{\mathcal{M}} \left(\begin{array}{l} P(U_1,\ldots,U_{n-1},V_1,\ldots,V_{n-1}) \\ \cdot\Big(f(V_1,\ldots,V_{n-1}) + \psi(U_{n-1},V_{n-1})\Big) \end{array} \right) .$$

As the next step, we will for each $v_{n-1} \in \mathcal{V}$ find a value u_{n-1} of decision variable U_{n-1} that maximizes E_f over the terms containing U_{n-1}. Note that the value of U_{n-1} cannot influence the past since when deciding on U_{n-1} the value of V_{n-1} is already known. It means that the values of V_{n-1} effectively separate the influence diagram into two parts and maximization over U_{n-1} can be performed only in the part containing U_{n-1}:

$$E_f^* = \max_{U_1,\ldots,U_{n-2}} \underset{V_1,\ldots,V_{n-1}}{\mathcal{M}} \left(\begin{array}{l} P(U_1,\ldots,U_{n-2},V_1,\ldots,V_{n-1}) \\ \cdot \max\limits_{U_{n-1}}\ \delta(U_{n-1}|V_{n-1}) \cdot \left(\begin{array}{l} f(V_1,\ldots,V_{n-1}) \\ +\psi(U_{n-1},V_{n-1}) \end{array} \right) \end{array} \right) .$$

Since $f(V_1,\ldots,V_{n-1})$ does not depend on U_{n-1} we get

$$E_f^* = \max_{U_1,\ldots,U_{n-2}} \underset{V_1,\ldots,V_{n-1}}{\mathcal{M}} \left(\begin{array}{l} P(U_1,\ldots,U_{n-2},V_1,\ldots,V_{n-1}) \\ \cdot(f(V_1,\ldots,V_{n-1}) + \psi(V_{n-1})) \end{array} \right) . \quad (28)$$

where

$$\psi(V_{n-1}) = \max_{U_{n-1}}\ \psi(U_{n-1},V_{n-1}) .$$

From formula (23) we can get formula (28) by substituting $n-1$ for n. Therefore we can repeat the transformations again and again until $n = 2$. In case $n = 2$ formula (28) reduces to

$$E_f^* = \underset{V_1}{\mathcal{M}}\ P(V_1) \cdot \psi(V_1) ,$$

which is formula (17) of the theorem we want to prove. □

Stochastic safety radius on Neighbor-Joining method and Balanced Minimal Evolution on small trees

Jing Xi
Department of Mathematics
North Carolina State University
jxi2@ncsu.edu

Jin Xie
Department of Statistics
University of Kentucky
jin.xie@uky.edu

Ruriko Yoshida
Department of Statistics
University of Kentucky
ruriko.yoshida@uky.edu

Stefan Forcey
Department of Mathematics
University of Akron
sforcey@uakron.edu

Abstract

A distance-based method to reconstruct a phylogenetic tree with n leaves takes a distance matrix, $n \times n$ symmetric matrix with 0s in the diagonal, as its input and reconstructs a tree with n leaves using tools in combinatorics. A safety radius is a radius from a tree metric (a distance matrix realizing a true tree) within which the input distance matrices must all lie in order to satisfy a precise combinatorial condition under which the distance-based method is guaranteed to return a correct tree. A stochastic safety radius is a safety radius under which the distance-based method is guaranteed to return a correct tree within a certain probability. In this paper we investigated stochastic safety radii for the neighbor-joining (NJ) method and balanced minimal evolution (BME) method for $n = 5$.

1 Introduction

A *phylogenetic tree* (or *phylogeny*) on the set $X = [n]$ is a graph which summarizes the relations of evolutionary descent between different species, organisms, or genes. Phylogenetic trees are useful tools for organizing many types of biological information, and for reasoning about events which may have occurred in the evolutionary history of an organism. There has been much research on phylogenetic tree reconstructions from alignments, and *distance-based* methods are some of the best-known phylogenetic tree reconstruction methods.

Once we compute pairwise distances $\forall(x, y) \in X \times X$ from an alignment, we can reconstruct a phylogenetic tree via distance-based methods. In contrast with parsi-

mony methods, distance-based methods have been shown to be statistically consistent in all settings (such as the long branch attraction) [7, 3, 4, 1]. Distance-based methods also have a huge speed advantage over parsimony and likelihood methods in terms of computational time, and hence enable the reconstruction of trees with large numbers of taxa. However, a distance-based method is not a perfect method to reconstruct a phylogenetic tree from the input sequence data set: in the process of computing a pairwise distance, we ignore interior nodes of a tree as well as a tree topology, and thus we lose information from the input sequence data sets. Therefore it is important to understand how a distance based method works and how robust it is with noisy data sets.

One way to measure its robustness is called the safety radius. A safety radius is a radius from a tree metric (a distance matrix realizing a true tree) within which the input distance matrices must all lie in order to satisfy a precise combinatorial condition under which the distance-based method is guaranteed to return a correct tree. More precisely, we have the following definition.

Definition 1. *Suppose we have a vector representation of all pairwise distances* $\delta \in \mathbb{R}^{\binom{n}{2}}$ *and suppose* $d_{T,w} := (d_{xy})_{x,y\in X}$*, where* $T \in \tau_n$*,* τ_n *is the set of all phylogenetic unrooted trees with leaves* $X = [n]$*, and* $w \in \mathbb{R}_+^{2n-3}$*, where* $\mathbb{R}_+$ *is the set of all non-negative real numbers, is a vector representation of the set of branch lengths in* T*, is a tree metric, i.e.,* $d_{xy} \geq 0$ *is the total of branch lengths in the unique path from a leaf* x *to a leaf* y *in* T*. Let* $w_{\min}$ *be the smallest interior branch length in* T*. Then a method* M *for reconstructing a phylogenetic* X*-tree from each distance matrix* δ *on* X *is said to have a* l_∞ *safety radius* ρ_n *if for any binary phylogenetic tree* T *with* n *leaves we have:*

$$||\delta - d_{T,w}||_\infty < \rho_n \cdot w_{\min} \Rightarrow M(\delta) = T.$$

Notice that the definition of the safety radius defined in Definition 1 is deterministic even though the input data δ is a multivariate random variable. Thus, this is more meaningful to define in terms of probability distribution. Thus, in 2014 Steel and Gascuel introduced a notion of *stochastic safety radius* [9].

Definition 2 (Stochastic safety radius)**.** *Suppose we allow* σ^2 *to depend on* n*:* $\sigma^2 = \dfrac{c^2}{\log(n)}$*, for some value* $c \neq 0$*. For any* $\eta > 0$*, we say that a distance-based tree reconstruction method* M *has* η*-stochastic safety radius* $s = s_n$ *if for every binary phylogenetic* X*-tree* T *on* n *leaves, with minimum interior edge length* $w_{\min}$*, and with the distance matrix* δ *on* X *described by the random errors model, we have*

$$c < s \cdot w_{\min} \Rightarrow P(M(\delta) = T) \geq 1 - \eta.$$

In this paper we focus on two distance-based methods, namely *neighbor-joining* (NJ) method and *balanced minimal evolution* (BME) method. In 2002, Desper and Gascuel introduced a BME principle, based on a branch length estimation scheme of Pauplin [13]. The guiding principle of minimum evolution tree reconstruction methods is to return a tree whose total length (sum of branch lengths) is minimal, given an input dissimilarity map. The BME method is a special case of these distance-based

methods wherein branch lengths are estimated by a weighted least-squares method (in terms of the input δ and the tree $T \in \tau_n$ in question) that puts more emphasis on shorter distances than longer ones. Each labeled tree topology gives rise to a vector, called herein *the BME vector,* which is obtained from Pauplin's formula. In 2000, Pauplin showed that the BME method is equivalent to optimizing a linear function, the dissimilarity map, over the BME representations of binary trees, given by the BME vectors [13]. Eickmeyer et. al. defined the n^{th} *BME polytope* as the convex hull of the BME vectors for all binary trees on a fixed number n of taxa. Hence the BME method is equivalent to optimizing a linear function, namely, the input distance matrix δ, over a BME polytope. They characterized the behavior of the BME phylogenetics on such data sets using the BME polytopes and the *BME cones*, i.e., the normal cones of the BME polytope.

The study of related geometric structures, the BME cones, further clarifies the nature of the link between phylogenetic tree reconstruction using the BME criterion and using the NJ Algorithm developed by Saitou and Nei [14]. In 2006, Gascuel and Steel showed that the NJ Algorithm, one of the most popular phylogenetic tree reconstruction algorithms, is a greedy algorithm for finding the BME tree associated to a distance matrix δ [8]. The NJ Algorithm relies on a particular criterion for iteratively selecting cherries; details on cherry-picking and the NJ Algorithm are recalled later in the paper. In 2008, based on the fact that the selection criterion for cherry-picking is linear in the distance matrix δ [2], Eickmeyer et. al. showed that the NJ Algorithm will pick cherries to merge in a particular order and output a particular tree topology T if and only if the pairwise distances satisfy a system of linear inequalities, whose solution set forms a polyhedral cone in $\mathbb{R}^{\binom{n}{2}}$ [5]. They defined such a cone as an *NJ cone*. In general, the sequence of cherries chosen by the NJ Algorithm is not unique, hence multiple distance matrix δ will be assigned by the NJ Algorithm to a single fixed tree topology T. The set of all distance matrix δ for which the NJ Algorithm returns a fixed tree topology T is a union of NJ cones, however this union is not convex in general. Eickmeyer et. al. characterized those dissimilarity maps for which the NJ Algorithm returns the BME tree, by comparing the NJ cones with the BME cones, for eight or fewer taxa [5].

In this paper we use the BME cones and NJ cones in order to investigate their stochastic safety radius for $n = 5$. Here we assume that the multivariate random variable δ is defined as follows:

$$\delta_{xy} = d_{xy} + \epsilon_{xy},$$

where $\epsilon_{xy} \sim N(0, \sigma^2)$, the Gaussian distribution with mean 0 and a standard deviation $\sigma > 0$, are independent for all pairwise distance $(x, y) \in X \times X$. This paper is organized as follows: Section 2 shows the probability distribution of a random δ so that it satisfies the *four point rule* for all distinct quartets in $[n]$ with a fixed T. Zarestkii in [11] defined the notion of the four point rule as follows: we select the tree topology $xy|wz$ (which means there is an internal edge between $\{x, y\}$ and $\{w, z\}$ for a distinct $x, y, w, z \in [n]$) if

$$\delta_{xy} + \delta_{wz} < \min\{\delta_{xw} + \delta_{yz},\, \delta_{xz} + \delta_{yw}\}. \tag{1}$$

In Section 3 we will show multivariate probability distribution $P(M(\delta) = T)$ where $T \in \tau_5$ is fixed and M is the BME method, in Section 4 shows the multivariate probability distribution $P(M(\delta) = T)$ where $T \in \tau_5$ is fixed and M is the NJ method. Finally in Section 5 we will show some computational results on these probability distributions and we have shown the plot for the stochastic safety radii for the NJ and the BME methods varying η and c for $n = 5$ (Figure 6). As shown in Figure 6 both stochastic safety radii are basically almost identical in this case since the probability distributions $P(M(\delta) = T)$ for the NJ and for the BME methods are almost identically same shown in Figure 5 for $n = 5$ and $w_{\min} = 1$.

2 Probability distribution on "four point rule"

For a tree containing random errors, the pairwise distance between two leaves is

$$\delta_{xy} = d_{xy} + \epsilon_{xy}$$

where x and y are different taxas of a tree, d_{xy} is the true pairwise distance between taxa x and y, and $\epsilon'_{xy}s$ follow i.i.d. Gaussian Distribution with mean 0 and variance σ^2. Intuitively in this section we are computing a probability distribution such that if we select a random $\delta \in \mathbb{R}^{\binom{n}{2}}$, δ satisfies Equation 1 if and only if there is an internal edge between $\{x, y\}$ and $\{w, z\}$ in $T \in \tau_n$, for all distinct $\{x, y, w, z\} \in [n]$. We find a formula for the probability, for 5 taxa, that a tree metric with random errors still obeys the original four-point inequalities on each subset of four leaves.

We first consider four point rule on 4 taxa tree. Suppose Figure 1 is the true tree. Then for a random tree, the following inequalities must be satisfied in order to return the correct tree:

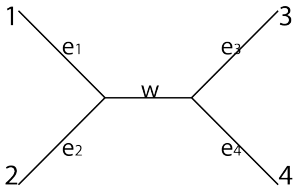


Figure 1: 4 taxa tree

$$\begin{array}{l} \delta_{12} + \delta_{34} \leq \delta_{13} + \delta_{24} \\ \delta_{12} + \delta_{34} \leq \delta_{14} + \delta_{23} \end{array} \quad (2)$$

Since

$$\begin{array}{rcl} \delta_{12} & = & e_1 + e_2 + \epsilon_{12} \\ \delta_{34} & = & e_3 + e_4 + \epsilon_{34} \\ \delta_{13} & = & e_1 + e_3 + w + \epsilon_{13} \\ \delta_{24} & = & e_2 + e_4 + w + \epsilon_{24} \\ \delta_{14} & = & e_1 + e_4 + w + \epsilon_{14} \\ \delta_{23} & = & e_2 + e_3 + w + \epsilon_{23} \end{array} \quad (3)$$

Then we can have

$$\begin{array}{l} \epsilon_{12} + \epsilon_{34} \leq 2w + \epsilon_{13} + \epsilon_{24} \\ \epsilon_{12} + \epsilon_{34} \leq 2w + \epsilon_{14} + \epsilon_{23} \end{array} \quad (4)$$

Since $\epsilon_{xy} \overset{iid}{\sim} N(0, \sigma^2)$, we know $\epsilon_{12} + \epsilon_{34}, \epsilon_{13} + \epsilon_{24}, \epsilon_{14} + \epsilon_{23} \overset{iid}{\sim} N(0, 2\sigma^2)$. Let f and F be the density and cumulative distribution functions of $N(0, 1)$, respectively. Then the probability that Inequality 4 is satisfied, i.e. the probability that a random distance matrix δ returns the true tree, equals to:

$$\int_{-\infty}^{\infty} f(x)[1 - F(\frac{x - 2w}{\sqrt{2}\sigma})]^2 dx \tag{5}$$

Now we consider four point rule on 5 taxa tree. Suppose the true tree is Figure 2(a). We need to check the rule on all possible combinations of four distinct leaves in this tree. It is trivial to see we only have 5 different combinations. For each of them, we could construct two inequalities similar to the way we obtained Equation 4. Therefore, we have 10 inequalities for the 5 combinations of 4 distinct taxa:

$$\begin{array}{rcl}
\epsilon_{12} + \epsilon_{34} & \leq & 2w_1 + \epsilon_{13} + \epsilon_{24} \\
\epsilon_{12} + \epsilon_{34} & \leq & 2w_1 + \epsilon_{14} + \epsilon_{23} \\
\epsilon_{12} + \epsilon_{35} & \leq & 2w_1 + \epsilon_{13} + \epsilon_{25} \\
\epsilon_{12} + \epsilon_{35} & \leq & 2w_1 + \epsilon_{15} + \epsilon_{23} \\
\epsilon_{12} + \epsilon_{45} & \leq & 2w_1 + 2w_2 + \epsilon_{14} + \epsilon_{25} \\
\epsilon_{12} + \epsilon_{45} & \leq & 2w_1 + 2w_2 + \epsilon_{15} + \epsilon_{24} \\
\epsilon_{13} + \epsilon_{45} & \leq & 2w_2 + \epsilon_{14} + \epsilon_{35} \\
\epsilon_{13} + \epsilon_{45} & \leq & 2w_2 + \epsilon_{15} + \epsilon_{34} \\
\epsilon_{23} + \epsilon_{45} & \leq & 2w_2 + \epsilon_{24} + \epsilon_{35} \\
\epsilon_{23} + \epsilon_{45} & \leq & 2w_2 + \epsilon_{25} + \epsilon_{34}
\end{array} \tag{6}$$

Let $\boldsymbol{\epsilon} \equiv (\epsilon_{12}, \epsilon_{13}, \epsilon_{14}, \ldots, \epsilon_{45})^T_{10\times 1}$ and

$$U = \begin{pmatrix}
1 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\
1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\
1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\
0 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1
\end{pmatrix}_{10\times 10},$$

then the 10 inequalities are:

$$U\boldsymbol{\epsilon} \leq (2w_1, 2w_1, 2w_1, 2w_1, 2w_1 + 2w_2, 2w_1 + 2w_2, 2w_2, 2w_2, 2w_2, 2w_2)^T. \tag{7}$$

Thus the probability that a 5-leaved tree metric with random errors still obeys the original-four point inequalities on each subset of four leaves is the probability that inequality (7) is satisfied.

3 Probability distribution of the output tree via the BME method

This method begins with a given set of n items and a symmetric (or upper triangular) square $n \times n$ *distance matrix* whose entries are numerical dissimilarities, or distances, between pairs of items. From the distance matrix the BME method constructs a binary tree with the n items labeling the n leaves. The BME tree has the property that the distances between its leaves most closely match the given distances between corresponding pairs of taxa.

By "most closely match" in the previous paragraph we mean the following: the reciprocals of the distances between leaves are the components of a vector $\mathbf{c}$, and this vector minimizes the dot product $\mathbf{c} \cdot \delta$ where δ is the list of distances in the upper triangle of the distance matrix.

More precisely: Let the set of n distinct species, or taxa, be called X. For convenience we will often let $X = [n] = \{1, 2, \ldots, n\}$. Let vector δ be given, having $\binom{n}{2}$ real valued components δ_{xy}, one for each pair $\{x, y\} \subset X$. There is a vector $\mathbf{c}(t)$ for each binary tree t on leaves X, also having $\binom{n}{2}$ components $c_{xy}(t)$, one for each pair $\{x, y\} \subset X$. These components are ordered in the same way for both vectors, and we will use the lexicographic ordering: $\delta = (\delta_{12}, \delta_{13}, \ldots, \delta_{1n}, \delta_{23}, \delta_{24}, \ldots, \delta_{n-1,n})$.

We define, following Pauplin [13]:

$$\mathbf{c}_{xy}(t) = \frac{1}{2^{l(x,y)}}$$

where $l(x, y)$ is the number of internal nodes (degree 3 vertices) in the path from leaf x to leaf y. The BME tree for the vector δ is the binary tree t that minimizes $\delta \cdot \mathbf{c}(t)$ for all binary trees on leaves X. Rather than the original fractional coordinates $\mathbf{c}_{xy}$ we will scale by a factor of 2^{n-2}, giving coordinates

$$\mathbf{x}_{xy} = 2^{n-2}\mathbf{c}_{xy} = 2^{n-2-l(x,y)}.$$

Since the furthest apart any two leaves may be is a distance of $n - 2$ internal nodes, this scaling will result in integral coordinates. Thus we can equivalently say that the BME tree for the vector δ is the binary tree t that minimizes $\delta \cdot \mathbf{x}(t)$ for all binary trees on leaves X.

Consider a tree metric d_T which arises from a binary tree T with five leaves $\{a, b, c, d, e\}$. Let the interior edges e_1 and e_2 have lengths $w_i = l(e_i)$.

Theorem 3. *Let T, the tree for which d_T is a tree metric, have cherries $\{a, b\}$ and $\{c, d\}$.*

Let:

$y_1 = 2\epsilon_{ac} + \epsilon_{ad} - \epsilon_{bc} + \epsilon_{bd} + 3\epsilon_{be} - \epsilon_{ce}$
$y_2 = 2\epsilon_{ac} + \epsilon_{ae} - \epsilon_{bc} + 3\epsilon_{bd} + \epsilon_{be} - \epsilon_{cd}$
$y_3 = -\epsilon_{ac} + \epsilon_{ad} + 3\epsilon_{ae} + 2\epsilon_{bc} + \epsilon_{bd} - 1\epsilon_{ce}$
$y_4 = -\epsilon_{ac} + 3\epsilon_{ad} + \epsilon_{ae} + 2\epsilon_{bc} + \epsilon_{be} - \epsilon_{cd}$

Let:

$$\begin{aligned} z_1 &= -\epsilon_{ac} + 3\epsilon_{ad} + \epsilon_{bd} + \epsilon_{be} - \epsilon_{cd} + 2\epsilon_{ce} \\ z_2 &= -\epsilon_{ac} + 3\epsilon_{ae} + \epsilon_{bd} + \epsilon_{be} + 2\epsilon_{cd} - \epsilon_{ce} \\ z_3 &= \epsilon_{ad} + \epsilon_{ae} - \epsilon_{bc} + 3\epsilon_{bd} - \epsilon_{cd} + 2\epsilon_{ce} \\ z_4 &= \epsilon_{ad} + \epsilon_{ae} - \epsilon_{bc} + 3\epsilon_{be} + 2\epsilon_{cd} - \epsilon_{ce} \end{aligned}$$

Then the BME method will return the correct tree T if and only if:

$$\begin{aligned} 4w_2 &> \epsilon_{ac} + \epsilon_{bc} + 2\epsilon_{de} - \min(\epsilon_{ad} + \epsilon_{bd} + 2\epsilon_{ce},\ \epsilon_{ae} + \epsilon_{be} + 2\epsilon_{cd}) \\ 6w_1 + 4w_2 &> 3\epsilon_{ab} + 2\epsilon_{de} - \min(y_1,\ y_2,\ y_3,\ y_4) \\ 6w_1 + 6w_2 &> 3\epsilon_{ab} + 3\epsilon_{de} - \min(3\epsilon_{ae} + 3\epsilon_{bd},\ 3\epsilon_{ad} + 3\epsilon_{be}) \\ 4w_1 + 6w_2 &> 2\epsilon_{ab} + 3\epsilon_{de} - \min(z_1,\ z_2,\ z_3,\ z_4) \\ 4w_1 &> 2\epsilon_{ab} + \epsilon_{cd} + \epsilon_{ce} - \min(\epsilon_{ad} + \epsilon_{ae} + 2\epsilon_{bc},\ 2\epsilon_{ac} + \epsilon_{bd} + \epsilon_{be}) \end{aligned}$$

Proof. The BME method will return the correct tree T if and only if

$$(d_T + \epsilon) \cdot \mathbf{x}(T) < (d_T + \epsilon) \cdot \mathbf{x}(t)$$

for all alternate trees t. This is true since the 1-skeleton of the BME polytope for $n = 5$ is the complete graph on the 15 vertices.

Further, the above inequality holds iff

$$d_T \cdot (\mathbf{x}(t) - \mathbf{x}(T)) > \epsilon \cdot (\mathbf{x}(T) - \mathbf{x}(t))$$

for all alternate trees t. Note that all the trees with five leaves have the same topology.

There are 14 other possible trees t. These separate into 5 sets of trees for which the left hand side of the above inequality is respectively $4w_2, 6w_1 + 4w_2, 6w_1 + 6w_2, 4w_1 + 6w_2$, or $4w_1$. For each of these we collect the right hand sides, and take their maximum. □

4 Probability distribution of the output tree via the NJ method

4.1 H-representation of NJ cones [6]

Recall that the tree metric $d_{T,w} = (d_{xy})_{1 \le x,y \le n}$ is a symmetric matrix with $d_{xx} = 0$. We can flatten the entries in the upper triangle (diagonal entries are omitted) by columns:

$$d_{xy} = d_{\frac{(y-1)(y-2)}{2} + x},$$

where $1 \le x \le n-1$ and $x+1 \le y \le n$. Let $\mathbf{d} = (d_1, d_2, \ldots, d_m)$, $m := \binom{n}{2}$, be the vector of tree metric after flattening. Notice here this flattening defines a one-to-one mapping between the indices:

$$I_f : \{(x,y) \in \mathbb{Z} : 1 \le x \le n-1, x+1 \le y \le n\} \to \{1,2,\ldots,m\},\ I_f(x,y) = \frac{(y-1)(y-2)}{2} + x.$$

In NJ algorithm, we first compute the Q-criterion (cherry picking criterion):

$$q_{xy} = (n-2)d_{xy} - \sum_{z=1}^{n} d_{xz} - \sum_{z=1}^{n} d_{zy}.$$

Similar as the flattened tree metric $\mathbf{d}$, the Q-criterion can also be flattened to a m dimensional vector $\mathbf{q}$ which can be obtained from $\mathbf{d}$ by linear transformation:

$$\mathbf{q} = A^{(n)}\mathbf{d},$$

where matrix $A^{(n)}$ is defined as:

$$A^{(n)}_{ij} = \begin{cases} n-4 & \text{if } i = j \\ -1 & \text{if } i \ne j \text{ and } \{x,y\} \cap \{z,w\} \ne \emptyset \\ 0 & \text{else} \end{cases},$$

where $(x,y) = I_f^{-1}(i)$ and $(z,w) = I_f^{-1}(j)$.

Now each entry in $\mathbf{q}$ corresponds to a pair of nodes in T, the next step of NJ algorithm is to find the minimum entry of $\mathbf{q}$ and join the corresponding two nodes as a cherry ("cherry picking"), then these two nodes will be replaced by a new node and the tree metric will be updated as $\mathbf{d}'$ (the dimension is reduced). We can see NJ algorithm proceeds by picking one cherry and reducing the size of the tree metric in each iteration until a binary tree is reconstructed. Without loss of generality and for the convenience of expression, we will only give details for the first iteration and assume the cherry we pick is $(n-1,n)$ in the rest part of this section.

First, to make cherry $(n-1,n)$ be the one to be picked, $q_m = q_{I_f(n-1,n)}$ needs to be the minimum in $\mathbf{q}$. This means the following inequalities need to be satisfied:

$$(I_{m-1}, -\mathbf{1}_{m-1})\mathbf{q} \ge 0 \Longrightarrow H^{(n)}\mathbf{d} \ge 0,\ H^{(n)} = (I_{m-1}, -\mathbf{1}_{m-1})A^{(n)}.$$

Note that if an arbitrary cherry is picked, then a permutation of columns need to be assigned to $H^{(n)}$.

Then, after picking cherry $(n-1,n)$, we join these two nodes as the new node $(n-1)^*$. Again, we can produce the new reduced tree metric from $\mathbf{d}$ by linear transformation:

$$\mathbf{d}' = R\mathbf{d},$$

where $R = (r_{ij}) \in \mathbb{R}^{(m-n+1)\times m}$,

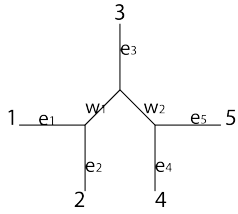
$$r_{ij} = \begin{cases} 1 & \text{if } 1 \le i = j \le \binom{n-2}{2} \\ 1/2 & \text{if } \binom{n-2}{2} + 1 \le i \le \binom{n-1}{2},\ j = i \\ 1/2 & \text{if } \binom{n-2}{2} + 1 \le i \le \binom{n-1}{2},\ j = i+n-2 \\ -1/2 & \text{if } \binom{n-2}{2} + 1 \le i \le \binom{n-1}{2},\ j = m \\ 0 & \text{else} \end{cases}.$$

At last, after including inequalities in all iterations, by the shifting lemma in [6], we also include the following equalities: $\forall$ node x,

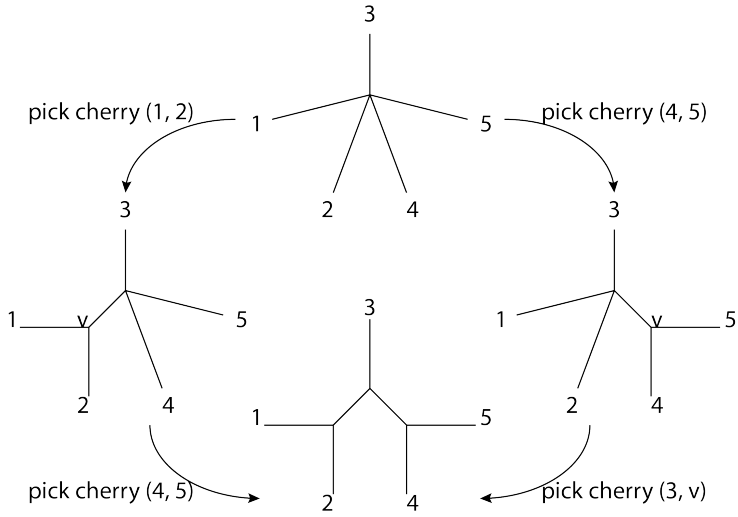
$$s_x^T \mathbf{d} = 0,\ (s_x)_i = \begin{cases} 1 & \text{if } x \in I_f^{-1}(i) \\ 0 & \text{else} \end{cases}.$$

4.2 H-representation of 5 taxa NJ cones

There is only one tree topology for 5 taxa tree. Therefore, without loss of generality, we assume our true tree is Figure 2(a).



(a) The 5 taxa tree



(b) Two orderings of picking cherries for 5 taxa tree

Figure 2: The 5 taxa tree used to generate data, all edges has length 1

For 5 taxa tree, the flattening for tree metric is:

$$\begin{array}{cccccccccccc} \text{xy:} & 12 & 13 & 23 & 14 & 24 & 34 & 15 & 25 & 35 & 45 \\ \mathbf{d} = & (d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10}) \end{array}$$

In Section 4.1, we can see that the permutation of columns for $H^{(n)}$ and R depends on the cherry we pick. This means that we should compute NJ cones for all ordering of cherry picking (see the two orderings of cherry picking in Figure 2(b) for example).

There are four orderings of cherry picking. First we can pick cherry $(1,2)$ then pick cherry $(4,5)$, which we denote as $(1,2)-(4,5)$. Use a similar notation we have the other three: $(1,2)-((1,2),3)$, $(4,5)-(1,2)$, and $(4,5)-(3,(4,5))$. Take the ordering $(4,5)-(3,(4,5))$ for example, use the results in Section 4.1 we can obtain the following linear constraints on $\mathbf{d}$:

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 1 & -1 \\ -1 & 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & -1 \\ -1 & -1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & -1 \\ -1 & -1 & 0 & 2 & 0 & 0 & 0 & 1 & 1 & -2 \\ -1 & 0 & -1 & 0 & 2 & 0 & 1 & 0 & 1 & -2 \\ 0 & -1 & -1 & 0 & 0 & 2 & 1 & 1 & 0 & -2 \\ -1 & -1 & 0 & 0 & 1 & 1 & 2 & 0 & 0 & -2 \\ -1 & 0 & -1 & 1 & 0 & 1 & 0 & 2 & 0 & -2 \\ 0 & -1 & -1 & 1 & 1 & 0 & 0 & 0 & 2 & -2 \\ -1 & 1 & 0 & 0 & 0.5 & -0.5 & 0 & 0.5 & -0.5 & 0 \\ -1 & 0 & 1 & 0.5 & 0 & -0.5 & 0.5 & 0 & -0.5 & 0 \end{pmatrix} \mathbf{d} \geq 0;$$

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \mathbf{d} = 0;$$

Although it is not obvious to see, we found that the linear constraints for orderings $(1,2)-(4,5)$ and $(1,2)-((1,2),3)$ are exactly the same, and the linear constraints for orderings $(4,5)-(1,2)$ and $(4,5)-(3,(4,5))$ are exactly the same. Therefore we only consider two NJ cones: the one for $(1,2)-(4,5)$ (denote as $\mathbf{C}_{(1,2)-(4,5)}$), and the one for $(4,5)-(1,2)$ (denote as $\mathbf{C}_{(4,5)-(1,2)}$).

4.3 Computing the probability that NJ reconstructs the correct 5 taxa tree

For the 5 taxa tree given in Figure 2(a) under the random errors model, we know that the flattened δ should follow a multi-variate normal (MVN) distribution with mean $\mu = (2,3,3,4,4,3,4,4,3,2)$ and covariance matrix $\Sigma = \sigma^2 I_{10}$. To trace the performance of NJ algorithm under different variation, we let σ^2 to be a set of values in $(0,1]$ and then compute the probability that NJ algorithm reconstructs the correct tree for each value of σ^2.

For a given σ^2, it is trivial to see that the probability that NJ algorithm returns the right tree is $Pr(\delta \in \mathbf{C}_{(1,2)-(4,5)}) + Pr(\delta \in \mathbf{C}_{(4,5)-(1,2)}) - Pr(\delta \in \mathbf{C}_{(1,2)-(4,5)} \cap \mathbf{C}_{(4,5)-(1,2)})$.

We used software **Polymake** [10] and verified that the dimension of $\mathbf{C}_{(1,2)-(4,5)} \cap \mathbf{C}_{(4,5)-(1,2)}$ is lower than both of them, therefore $Pr(\delta \in \mathbf{C}_{(1,2)-(4,5)} \cap \mathbf{C}_{(4,5)-(1,2)}) = 0$. **Polymake** also gives us the facets of these two NJ cones. For example, the facets of $\mathbf{C}_{(4,5)-(1,2)}$ are:

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 1 & -1 \\ -1 & -1 & 0 & 2 & 0 & 0 & 0 & 1 & 1 & -2 \\ -1 & 0 & -1 & 0 & 2 & 0 & 1 & 0 & 1 & -2 \\ 0 & -1 & -1 & 0 & 0 & 2 & 1 & 1 & 0 & -2 \\ -1 & -1 & 0 & 0 & 1 & 1 & 2 & 0 & 0 & -2 \\ -1 & 0 & -1 & 1 & 0 & 1 & 0 & 2 & 0 & -2 \\ 0 & -1 & -1 & 1 & 1 & 0 & 0 & 0 & 2 & -2 \\ -1 & 0 & 1 & 0.5 & 0 & -0.5 & 0.5 & 0 & -0.5 & 0 \\ -1 & 1 & 0 & 0 & 0.5 & -0.5 & 0 & 0.5 & -0.5 & 0 \end{pmatrix} \mathbf{d} \geq 0;$$

With these facets, we can use the **R** function "pmvnorm{mvtnorm}" with GenzBretz algorithm to compute $Pr(\delta \in \mathbf{C}_{(1,2)-(4,5)})$ and $Pr(\delta \in \mathbf{C}_{(4,5)-(1,2)})$.

5 Computational experiments

In this section, we show both the theoretical and simulation probabilities that the four point rule reconstructs the correct tree, as well as NJ algorithm and BME method. In our computational experiments, we set all branch lengths to be 1's and compute the probabilities for different values of σ^2.

In Figure 3, when σ^2 is increasing, the probability of 5 taxa tree will dramatically decrease faster than the probability of 4 taxa tree because we have more constraints to satisfy in 5 taxa tree which leads to lower probabilities.

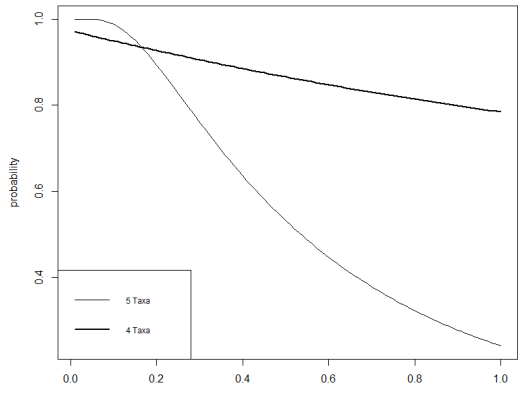
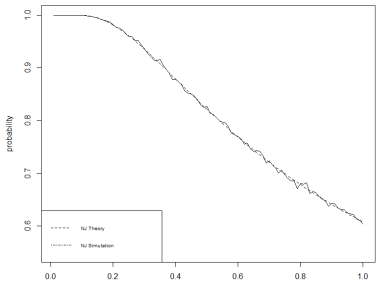
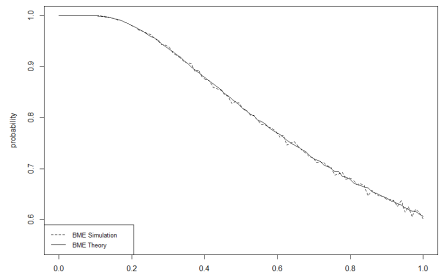


Figure 3: Theoretical probability that four point rule will return the correct 4 taxa tree, and that 5 taxa tree metric with random errors still obeys the original four-point inequalities on each subset of four leaves.

In Figure 4(a), we calculated the theoretical probability that the NJ method reconstructs the correct 5 taxa tree based on Section 4. For the simulation, we fix the 5 taxa tree in Figure 2(a) with all branch lengths to be $1's$, and add i.i.d. normal random errors $\epsilon'_{xy}s$ to the pairwise distance matrix. Then we use **R** function "nj{ape}" from the "ape" package in **R** [12] to reconstruct the tree. If the RF distance equals 0, it means that NJ method successfully returns the correct tree. Our simulation is based on 10,000 random trees. Figure 4(a) shows that the theoretical probabilities perfectly match the simulation result.



(a) Theoretical Probability and Simulation for NJ on 5 Taxa Tree



(b) Theoretical Probability and Simulation for BME on 5 Taxa Tree

Figure 4: Probability distributions for five leaves

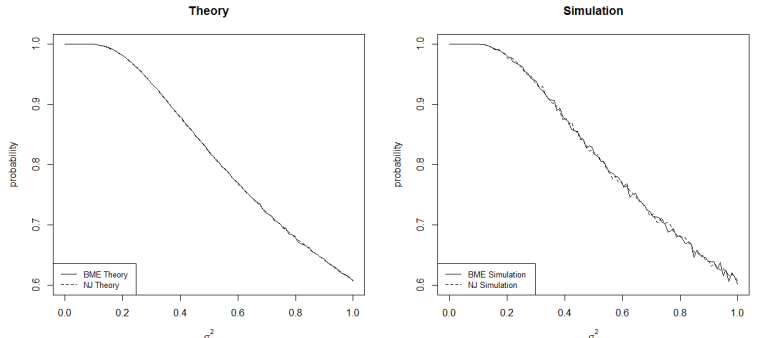


Figure 5: Comparison between BME and NJ on 5 Taxa Tree

In Figure 4(b), we calculated the theoretical and simulated probabilities that the BME method will return the correct 5 taxa tree. For the theoretical probability, we generate 100,000 sets of random errors, and check whether the theoretical rule is satisfied. In the end, we return the percentage. For the simulation, we generate random trees in the similar way to what we did for NJ algorithm. Then we used **R** function "fastme.bal{ape}" to reconstruct the tree. Again we used RF distance to check if the BME method successfully returned the correct tree. Our simulation is based on 10,000 random trees. Figure 4(b) shows that the theoretical probabilities perfectly match the simulation result.

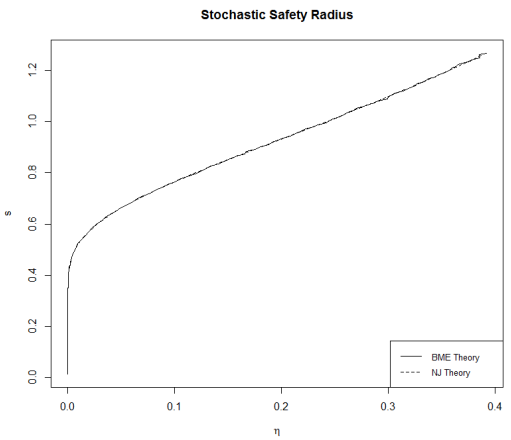


Figure 6: Stochastic safety radii for the NJ and BME methods for $n = 5$. The x-axis represents η and the y-axis represents the upper bound for $c/w_{\min}$ with $w_{\min} = 1$ for this experiment.

Figure 5 shows that there is almost no difference between BME and NJ methods on 5 taxa tree in both theoretical probabilities and simulation result.

Figure 6 shows the stochastic safety radii for the NJ and the BME methods for $n = 5$ and $w_{\min} = 1$. As shown in Figure 6 both stochastic safety radii are basically almost identical in this case since the probability distributions $P(M(\delta) = T)$ for the NJ and for the BME methods are almost identically same shown in Figure 5.

Acknowledgements

Stefan Forcey would like to thank the American Mathematical Society and the Mathematical Sciences Program of the National Security Agency for supporting this research through grant H98230-14-0121.[1]

References

[1] M. Bordewich, O. Gascuel, K. T. Huber, and V. Moulton. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 6(1):110–117, 2009.

[2] D Bryant. On the uniqueness of the selection criterion in neighbor-joining. *J. Classif.*, 22:3–15, 2005.

[3] Ronald W. DeBry. The consistency of several phylogeny-inference methods under varying evolutionary rates. *Mol Biol Evol*, 9(3):537–551, 1992.

[4] F. Denis and O. Gascuel. On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Applied Mathematics*, 127(1):63–77, 2003.

[5] K. Eickmeyer, P. Huggins, L. Pachter, and R. Yoshida. On the optimality of the neighbor-joining algorithm. *Algorithms for Molecular Biology*, 3(5), 2008.

[6] Kord Eickmeyer and Ruriko Yoshida. R: Geometry of neighbor-joining algorithm for small trees. In *Proceedings of the third international conference on Algebraic Biology*, 2008.

[7] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 22:240–249, 1978.

[8] O Gascuel and M Steel. Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000, 2006.

[9] O. Gasquel and M. Steel. A 'stochastic safety radius' for distance-based tree reconstruction, 2014.

[1]This manuscript is submitted for publication with the understanding that the United States Government is authorized to reproduce and distribute reprints.

[10] E Gawrilow and M Joswig. polymake: an approach to modular software design in computational geometry. In *Proceedings of the 17th Annual Symposium on Computational Geometry*, pages 222–231. ACM, 2001. June 3-5, 2001, Medford, MA.

[11] Zarestkii K. Reconstructing a tree from the distances between its leaves (in russian). *Uspehi Mathematicheskikh Nauk*, 20:90–92, 1965.

[12] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[13] Y. Pauplin. Direct calculation of a tree length using a distance matrix. *J. Mol. Evol.*, 51:41–47, 2000.

[14] N Saitou and M Nei. The neighbor joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

On Linear Probabilistic Opinion Pooling Based on Kullback-Leibler Divergence

extended abstract

Vladimíra Sečkárová
Department of Adaptive Systems
Institute of Information Theory and Automation of the CAS
Pod Vodárenskou věží 4, Prague 8, CZ 182 08
e-mail: seckarov@utia.cas.cz

Abstract

In this contribution we focus on the finite collection of sources, providing their opinions about a hidden (stochastic) phenomenon, that is not directly observable. The assumption on obtaining opinions yields a decision making process commonly referred to as opinion pooling. Due to the complexity of the space of possible decisions we consider the probability distributions over this set rather than single values, exploited before, e.g., in [2]. The final decision (result of pooling) is then a combination of probability distributions provided by sources. Here, we in particular exploit the combination introduced in [4], assuming each source is cooperating and willing to share its opinion with others, but *selfishly* requires the combination to be close to its opinion. The summary of basic steps is given below.

1 Kullback-Leibler divergence based combination of sources' opinions

Let us have $s < \infty$ sources providing discrete probability distributions represented by probability mass functions (pmf) $\mathbf{p}_1, \ldots, \mathbf{p}_s$:

$$\mathbf{p}_j = (p_{j1}, \ldots, p_{jn}) : \quad p_{ji} > 0,\ \sum_{i=1}^{n} p_{ji} = 1,\ n < \infty,\ j = 1, \ldots, s. \tag{1}$$

By exploiting theory of the Bayesian decision making [3] we search for their combination as the estimator $\hat{\mathbf{q}}$ of an unknown pmf $\mathbf{q}$ minimizing the expected Kullback-Leibler divergence [1]:

$$\mathrm{E}_{\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)}\mathrm{KLD}(\mathbf{q}||\hat{\mathbf{q}}). \tag{2}$$

The minimizer of (2) is

$$\hat{\mathbf{q}} = \mathrm{E}_{\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)}[\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s]. \tag{3}$$

To obtain the conditional expectation in (3) the conditional probability density function (pdf) $\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)$ has to be specified. We formalize the notion of *selfishness* among sources by considering the following equality constraints:

$$\mathrm{E}_{\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)}[\mathrm{KLD}(\mathbf{p}_j||\mathbf{q})|\mathbf{p}_1,\ldots,\mathbf{p}_s] = \mathrm{E}_{\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)}[\mathrm{KLD}(\mathbf{p}_s||\mathbf{q})|\mathbf{p}_1,\ldots,\mathbf{p}_s], \tag{4}$$

$j = 1,\ldots,s-1$. Let $\mathcal{S}$ denote the set of all pdfs $\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)$ satisfying (4). We now exploit the minimum cross-entropy principle [5] and choose the conditional pdf $\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s) \in \mathcal{S}$ that solves:

$$\min_{\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)\in\mathcal{S}} \mathrm{KLD}(\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)||\pi_0(\mathbf{q})), \tag{5}$$

where $\pi_0(\mathbf{q})$ denotes the prior guess on the conditional pdf $\pi(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)$.

We choose the pdf of the Dirichlet distribution with parameters $\nu_{01},\ldots,\nu_{0n}$ as the prior guess $\pi_0(\mathbf{q})$ for its computationally advantageous properties. Then, the conditional pdf $\hat{\pi}(\mathbf{q}|\mathbf{p}_1,\ldots,\mathbf{p}_s)$ minimizing (5) is also the pdf of the Dirichlet distribution $Dir(\hat{\nu}_1,\ldots,\hat{\nu}_n)$. The values of its parameters $\hat{\nu}_1,\ldots,\hat{\nu}_n$ are expressed by the following formula:

$$\hat{\nu}_i = \nu_{0i} + \sum_{j=1}^{s} \lambda_j (p_{ji} - p_{si}), \quad i = 1,\ldots,n, \tag{6}$$

where λ_j are the Lagrange multipliers resulting from minimization of (5) with respect to $(s-1)$ equations in (4), and the combination (3) is

$$\hat{q}_i = \frac{\nu_{0i}}{\sum_{k=1}^{n}\nu_{0k}} + \sum_{j=1}^{s} \frac{\lambda_j}{\sum_{k=1}^{n}\nu_{0k}} (p_{ji} - p_{si}), \quad i = 1,\ldots,n. \tag{7}$$

Although the combination has been introduced earlier in [4], its properties have not received much attention. We next discuss the choice of prior parameters $\nu_{01},\ldots,\nu_{0n}$ and the changes in the value of the combination when we deal with duplicate opinions.

2 Properties of the combination

It is somewhat surprising that the equation (6) combines simultaneously both, the parameters of the Dirichlet distribution and pmfs $\mathbf{p}_1,\ldots,\mathbf{p}_s$. Pmfs provided by sources can be viewed as individual guess for $\nu_1,\ldots,\nu_n$ when $\sum_{k=1}^{n}\nu_k = \sum_{k=1}^{n}\nu_{0k} = 1$. By plugging this relation into (7) we obtain

$$\hat{q}_i = p_{0i} + \sum_{j=1}^{s-1} \lambda_j p_{ji} + \left(-\sum_{j=1}^{s-1}\lambda_j\right) p_{si}, \tag{8}$$

where prior pmf $(p_{01}, \ldots, p_{0n})$, generally $p_{0i} = \frac{\nu_{0i}}{\sum_{k=1}^{n} \nu_{0k}}$, coincides with $(\nu_{01}, \ldots, \nu_{0n})$, a part of $\hat{\mathbf{q}}$ induced by prior pdf prior pdf $\pi_0(\mathbf{q})$.

Remind that we focus on combining sources' (experts') opinions, where the prior information about the studied problem may not be available. For the prior guess on $(p_{01}, \ldots, p_{0n})$ one should then exploit provided pmfs $\mathbf{p}_1, \ldots, \mathbf{p}_s$. Based on the additive nature of the derived optimal estimator $\hat{q}$ and the considered relation between $(\nu_{01}, \ldots, \nu_{0n})$ and $(p_{01}, \ldots, p_{0n})$ in (8), we focus on the weighted linear combination of $\mathbf{p}_1, \ldots, \mathbf{p}_s$, e.g., arithmetic mean. Preferences can be assigned by delegated person or depend on other available information, e.g., sources' prior information about parameters of the Dirichlet distribution. The constraints (equality of the expected KL-divergences) should then be modified accordingly.

We next study how the value of the combination (7) changes with the duplicate data. Let us now have $s+1$ pmfs $\mathbf{p}_1, \ldots, \mathbf{p}_s, \mathbf{p}_{s+1}$ and for simplicity assume that $p_{s+1,i} = p_{s,i}$, $i = 1, \ldots, n$. Let $\lambda_1, \ldots, \lambda_s$ be the Lagrange multipliers related to s equality constraints in (4). Then, for a fixed prior pmf $\mathbf{p}_0$, the combination of $\mathbf{p}_1, \ldots, \mathbf{p}_s, \mathbf{p}_{s+1}$ coincides with combination evaluated with omission of $\mathbf{p}_{s+1}$ and unchanged $\mathbf{p}_0$:

$$\hat{q}_i = p_{0i} + \sum_{j=1}^{s-1} \lambda_j (p_{ji} - p_{si}) + \lambda_s (p_{si} - p_{si}). \tag{9}$$

The additivity property of combination (7) implies that if other s_1 sources gave the same pmf $\mathbf{p}_k$, then the coefficient of each source equals $\frac{\lambda_k}{s_1}$.

It may seem strange that repeated sources' opinion are not taken more "seriously", with a higher weight. This is consequence of the fact that individual sources are not qualified by a weight reflecting their reliability. When such a weighting will be introduced, the coincidence of opinions can be taken into account and distinguished from cheating by repetitions of the same opinion.

Conclusion and future work

In this contribution we focused on approach to combining sources' opinions described in [4]. This combination is of a conservative type and qualifies all repetitions as "cheating", prevents overweighing of such source. The analogue of (7), where the prior guess $\mathbf{p}_0$ as well as the constraints (4) are influenced by preferences among sources, is of interest in the future.

Keywords

combining discrete probability distributions, linear opinion pooling, minimum cross-entropy principle

Acknowledgement

This research has been supported by GAČR 13-13502S.

References

[1] Bernardo, J. M.: Expected information as expected utility. *Ann. Stat.*, vol. 7 (1979)

[2] Kárný, M., Guy, T. V., Bodini, A., Ruggeri, F.: Cooperation via sharing of probabilistic information. *Int. J. of Computational Intelligence Studies*, vol. 1 (2009)

[3] Savage, L. J.: The Foundations of Statistics. Dover Books on Mathematics Series (1972)

[4] Sečkárová, V.: Dynamic Parameter Estimation Based on Minimum Cross-Entropy Method For Combining Information Sources. *Pliska Studia Mathematica Bulgarica*, vol. 24 (2015)

[5] Shore, J. E. and Johnson, R. W.: Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory*, vol. 26 (1980)

Index of Authors