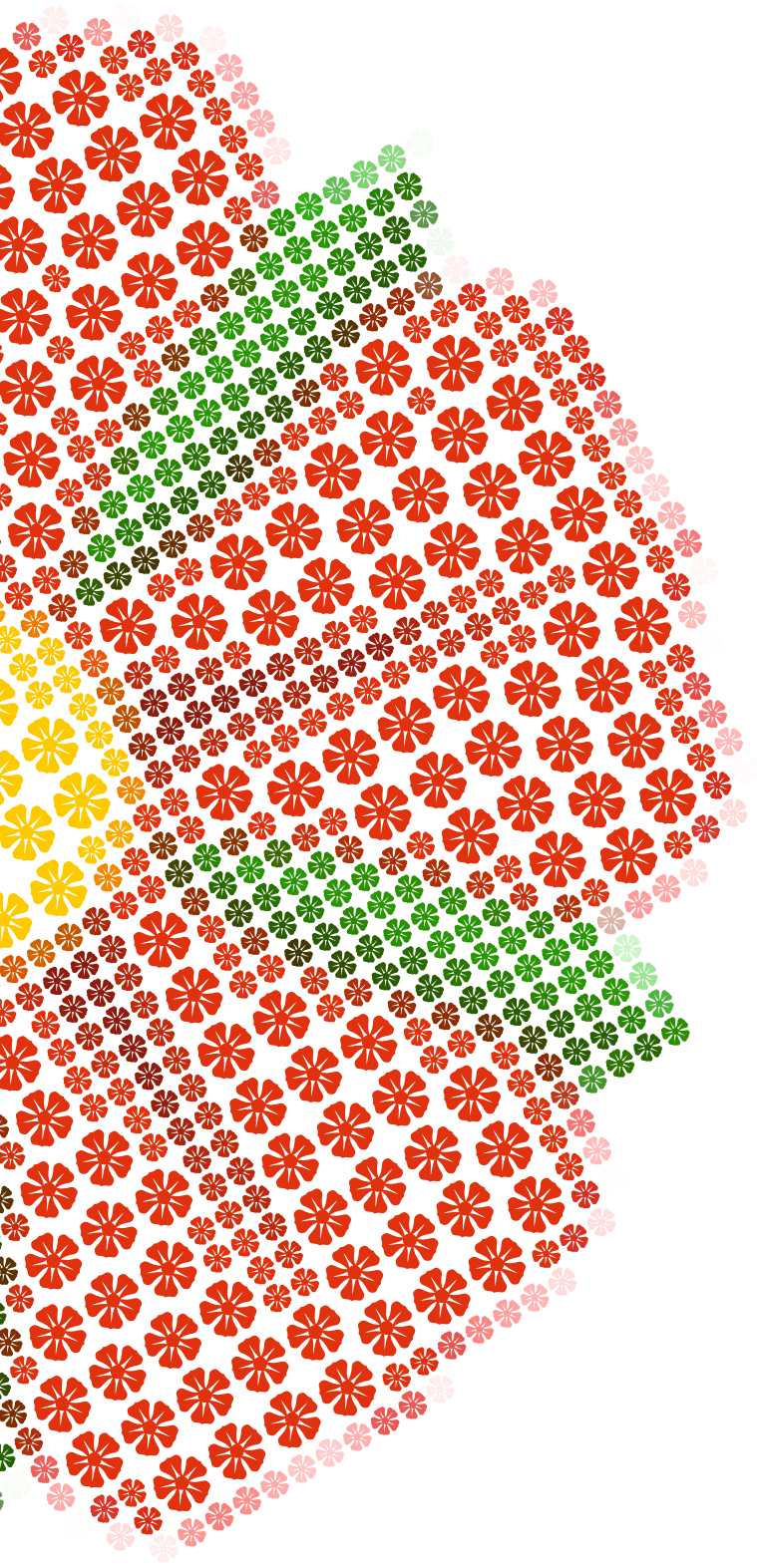


Proceedings of the 11th Workshop on Uncertainty Processing

WUPES'18

JUNE 6–9, 2018

TŘEBOŇ, CZECH REPUBLIC

Proceedings of the 11th Workshop on Uncertainty Processing *(WUPES'18)* Třeboň, Czech Republic

Václav Kratochvíl, Jiřina Vejnarová
(editors)

June 6-9, 2018

Published by:
MatfyzPress,
Publishing House of the Faculty of Mathematics and Physics Charles University
Sokolovská 83, 186 75 Praha 8, Czech Republic
as the 555. publication.
Printed by Reprostředisko MFF UK.

Printed in Prague — June 2018

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

Credits:
Editors: Václav Kratochvíl, Jiřina Vejnarová
LaTeXeditor: Václav Kratochvíl
Cover design: Jiří Přibil
using LaTeX's 'confproc' package, version 0.8 by V. Verfaille

ISBN: 978-80-7378-361-7

11th WORKSHOP ON UNCERTAINTY PROCESSING

WUPES '18
Třeboň

Organized by:

Institute of Information Theory and Automation,
Czech Academy of Sciences

&

Faculty of Management,
University of Economics, Prague

Třeboň

June 6-9, 2018

Programme and Conference Committee:
František Matúš - chair, *Czech Academy of Sciences, Prague*
Nihat Ay, *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*
Giulianella Coletti, *Università di Perugia, Italy*
Gernot D. Kleiter, *Universität Salzburg, Austria*
Prakash P. Shenoy, *University of Kansas, USA*

Organizing Committee:
Jirka Vomlel - chair, *Czech Academy of Sciences, Prague*
Radim Jiroušek, *Czech Academy of Sciences, Prague*
Václav Kratochvíl, *Czech Academy of Sciences, Prague*
Milan Studený, *Czech Academy of Sciences, Prague*
Jiřina Vejnarová, *Czech Academy of Sciences, Prague*

Foreword

The Workshop on Uncertainty Processing, better known under its abbreviation WUPES, celebrates its 30-year anniversary this year. In 1988, when the first Workshop took place, Czechoslovakia was still a communist country and a part of the Soviet bloc. Since then, many things have changed. For example, Czechoslovakia no longer exists as a country (because in 1993 it was peacefully split into two independent countries – Czechia and Slovakia). From this perspective, it is hard to believe that we have several participants who have attended most workshops in the the thirty-year history of WUPES.

As of now, the Program Committee has accepted, based on the extended abstracts, 21 papers to be presented at the Workshop, and 19 out of them are to be published in the present Conference Proceedings. These papers cover diverse topics, such as information processing, decision making, and data analysis; but what is common to most of them is that they are related to uncertainty calculus - Bayesian Networks, Dempster-Shafer Theory, Belief Functions, Probabilistic Logic, Game Theory, etc.

This Workshop takes place in Třeboň, a town in Southern Bohemia. Certainly, Třeboň is an interesting place with a rich history. I do not intend to make a review of Třeboň's history at this point; but I would like to mention its three moments directly related to our Workshop.

The Workshop logo contains a petal of a red wild rose. This is associated with the Coat of Arms of Rosenbergs, a significant and influential Czech noble family, which played an important role in the medieval history. The last reigning lord in this family was Petr Vok, whose residence was, guess where, in Třeboň, where he died in 1611.

The Workshop Dinner is going to take place in the restaurant named Kelly's Tavern. This restaurant is named after Edward Kelley, an English Renaissance occultist and self-declared spiritist medium. He worked with John Dee (an advisor to English Queen Elizabeth I) in his magical investigations. Besides the professed ability to summon spirits or angels, Kelley also claimed to possess the secret of transmuting base metals into gold, the true goal of alchemy, as well as the supposed Philosopher's Stone itself. But why is a tavern in a South Bohemian town named after Kelley? After an unsuccessful audience with Emperor Rudolf II in Prague Castle, Kelley and Dee found the patronage of the wealthy Bohemian Lord William of Rosenberg. Kelley and Dee settled in the town of Třeboň and continued their research there. According to Dee's diary, it was during that time that Kelley is said to have performed his first alchemical transmutation, namely, on December 19, 1586.

The Workshop lectures will be given in Schwarzenbergs' Hall of Třeboň Castle. The name of this hHall refers to another important noble family: Schwarzenberg is a Czech and German aristocratic family, and it was one of the most prominent European noble houses. The Schwarzenbergs achieved the rank of Princes of the Holy Roman Empire. From 1660, the House of Schwarzenberg was the owner of the Town and the Castle. During our Workshop walking trip, we will visit the Schwarzenberg Crypt in Domanín. The current head of the family is Karel, the 12th Prince of Schwarzenberg, a Czech politician who once served as the Minister of Foreign Affairs of the Czech Republic.

But back to the present. As is customary, this workshop is co-organized by two institutions – the Institute of Information Theory and Automation of the Czech Academy of Sciences, and the Faculty

of Management, University of Economics, Prague. We are grateful for their financial support. I would also like to thank all members of the Program Committee and of the Organizing Committee for their work, as well as to express my special gratitude to Radim Jiroušek, who was and still is the main person behind WUPES and has kept it running for thirty years! Another person who has helped a lot with organizing this Workshop is Václav Kratochvíl – many thanks!

I wish all participants pleasant participation in the Workshop with many interesting and stimulating discussions.

In Světice, May, 13, 2018

Jirka Vomlel

Information sources:

[1] Proceedings of the 10th Workshop on Uncertainty Processing (WUPES'15) Monínec, Czech Republic, Foreword,
http://wupes.fm.vse.cz/2015/data/Proceedings.pdf

[2] Rosenberg family, Wikipedia,
https://en.wikipedia.org/wiki/Rosenberg_family

[3] House of Schwarzenberg, Wikipedia,
https://en.wikipedia.org/wiki/House_of_Schwarzenberg

[4] Edward Kelley, Wikipedia,
https://en.wikipedia.org/wiki/Edward_Kelley

[5] John Dee, Wikipedia,
https://en.wikipedia.org/wiki/John_Dee

Fero Matúš

On Wednesday, May 16, 2018, Fero Matúš passed away. For all of us, it came as a shock. Yes, we had known that he was seriously ill; but still, Fero was a strong personality, full of new plans, and we thus were not prepared for this sudden end.

This bad piece of news has taken us unawares just when these Proceedings are ready for printing. We have not had time to elaborate on an official obituary surveying all of his scientific achievements. It would have been a long list, regardless of him being so young. He was so young that when the history of WUPES meetings was begun, Fero was still a university student. Nevertheless, he attended the second WUPES Conference in Alšovice in 1991, and he became a regular attendant at that time. We will strongly miss him this year. We will miss him all the more because it happens the first time in the 30-year history of WUPES Conferences that the chair of the Programme Committee does not take part at the meeting.

With his death, we are losing one of a few truly renaissance persons: he was a mathematician, musician, and gifted for languages. These Proceedings are thus, naturally, devoted to the memory of Fero Matúš. Needless to say, he will be present in our thoughts for the whole meeting in Třeboň. This meeting will be an opportunity for recalling his fine personal character. Also will it be an opportunity for recalling in what way he helped all of us, whether solutions to scientific problems, or by accepting unpleasant official positions (Head of Department or Member of the Scientific Board). But it will also be an opportunity to recall all nice real and often joyful stories we experienced together, and especially his typical saying with which he generously solved all the unpleasant and unimportant problems: *To sú také somariny* [What a hooey!].

On behalf of the Organizing Committee
Radim Jiroušek

List of Content

How to Generate the Network you Want with the PC Learning Algorithm

Emad Alsuwat
Computer Science and Engineering
University of South Carolina
Alsuwat@email.sc.edu

Marco Valtorta
Computer Science and Engineering
University of South Carolina
Mgv@cec.sc.edu

Csilla Farkas
Computer Science and Engineering
University of South Carolina
Farkas@cec.sc.edu

Abstract

Data integrity is a key component of effective Bayesian network structure learning algorithms, namely PC algorithm, design and use. Given the role that integrity of data plays in these outcomes, this research demonstrates the importance of data integrity as a key component in machine learning tools in order to emphasize the need for carefully considering data integrity during tool development and utilization. To meet this purpose, we study how an adversary could generate a desired network with the PC algorithm. Given a Bayesian network B_1 and a database DB_1 generated by B_1 and a second Bayesian network, B_2, which is equal to B_1, except for a minor change like a missing link, a reversed link, or an additional link, we explore and analyze what is the minimal number of changes such as additions, deletions, substitutions to DB_1 that lead to a database DB_2 that, when given as input to PC algorithm, results in B_2.
Keywords: Adversarial Machine Learning, Bayesian Networks, Data Poisoning Attacks, The PC Algorithm.

1 Introduction and Motivation

There has been a massive increase in the use of machine learning for diverse computer applications. Machine learning algorithms including Bayesian network algorithms are not secure against adversarial attacks. A machine learning algorithm is a *secure learning algorithm* if it functions well in adversarial environments [2]. It has been shown that an adversary can corrupt machine learning models by manipulating the input dataset [3, 24].

Therefore, it is essential to not only consider the presence of adversarial opponents but also design effective and resilient learning algorithms.

Adversarial machine learning is the research field that studies the design of efficient machine learning algorithms in adversarial environments [7]. In the presence of adversarial opponents (*the offense*), the problem of securing machine learning systems becomes harder since adversaries will attempt to corrupt machine learning models by crafting their inputs (known as *adversarial samples*) in an intelligent way. *Adversarial samples* are input datasets to machine learning algorithms that are crafted by adversaries to deliberately corrupt machine learning models [20]. Among different adversarial attacks, *data poisoning attacks*, which aim to corrupt the machine learning model by contaminating the data in the training phase, are considered one of the most important emerging security threats against machine learning systems [15].

Previous research has been conducted on data poisoning attacks against machine learning algorithms such as Support Vector Machines (SVMs) [3,5,9,14,17,25,26], Neural Networks (NNs) [27], and other machine learning algorithms [15]. Surprisingly, to date, no study has been performed on evaluating the vulnerabilities of Bayesian network learning algorithms against adversarial attacks.

In investigating data poisoning attacks on Bayesian network algorithms, we study attacks that aim to invalidate the Bayesian model either by modifying the input sample to delete the weakest edge or by adding cases to the original input sample to add the most believable yet incorrect edge.

In this paper, our main focus is to evaluate the robustness of Bayesian network learning algorithms on multiple adversarial samples against the proposed attacks. Our experiments show that the PC algorithm is vulnerable to data poisoning attacks. The learned Bayesian network model is vulnerable if an adversary can only inject a small number of adversarial samples into the original dataset.

We present our findings pertaining to the robustness of the PC algorithm against data poisoning attacks. The main contributions are the following: 1) We define a novel measure of the strengths of links between variables in Bayesian networks and present our detailed analysis. 2) We demonstrate how to use the defined link strengths measure to add edges to and delete edges from a Bayesian network model. 3) We present two types of data poisoning attacks against the PC structure learning algorithm and implement the attacks. 4) We have implemented our approach and presented the results.

The rest of this paper is structured as follows. In section 2, we present the background information. Our link strengths measure in discrete Bayesian networks is presented in section 3. In section 4, we present two types of data poisoning attacks against the PC algorithm. In section 5, we present our empirical results. In section 6, we provide conclusions and directions for future work.

2 Background Information

2.1 Structure Learning in Bayesian Networks

There are three main approaches to learning the structure of Bayesian networks: *constraint-based*, *score-based*, or *hybrid* algorithms. We will focus on constraint-based algorithms, namely the PC algorithm, since it is an integral part of this paper. *The PC algorithm* (named after the authors, the first letter of their first names, **P**eter Spirtes and **C**lark Glymour) is a constraint-based algorithm for learning the structure of a Bayesian network from data. The PC algorithm follows the theoretical framework of the IC algorithm to determine the structure of causal models [22]. According to [23], the process performed by the PC algorithm to learn the structure of Bayesian networks can be summarized as follows: (i) For every pair of variables, perform statistical tests for conditional independence. (ii) Determine the skeleton (undirected graph) of the learned structure by adding a link between every pair of statistically dependent variables. (iii) Identify colliders (v-structures) of the learned structure (A $\rightarrow$ B $\leftarrow$ C). (iv) Identify derived directions. (v) Randomly, complete orienting the remaining undirected edges without creating a new collider or a cycle. For the implementation of this paper, we used *the Hugin PC algorithm* (by $Hugin^{TM}$ *Decision Engine* [12, 19]), "which is a variant of the original PC algorithm due to [23]" [8].

2.2 Prior to Posterior Updating

Bayes' theorem is a simple mathematical formula that inverts conditional probabilities (i.e., given the conditional probability of event B given event A, how to calculate the conditional probability of event A given event B). The statement of Bayes' theorem is: For two events A and B, $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$, where (i) $P(A \mid B)$ is the conditional probability of event A given event B (called the posterior probability), (ii) $P(B \mid A)$ is the conditional probability of event B given event A (called the likelihood), (iii) $P(A)$ is the marginal probability of event A (called the prior probability), and (iv) $P(B)$ is the marginal probability of event B ($P(B) > 0$) [16].

Unlike classical statistics, Bayesian statistics treats parameters as random variables whereas data is treated as fixed. For Example, let θ be a parameter, and D be a dataset, then Bayes' theorem can be expressed mathematically as follows:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \tag{1}$$

In equation 1, $P(\theta \mid D)$ is the *posterior distribution*, which is the ultimate goal for Bayesian statistics since it measures the uncertainty about the parameters θ after seeing the dataset D. $P(D \mid \theta)$ is the *likelihood*, which describes how likely the dataset D is if the truth is parameter θ. $P(\theta)$ is the *prior distribution*, which is a marginal probability of our belief before seeing data. $P(D)$ is the *marginal probability* of D, which is a normalization constant to ensures that the sum of the posterior distribution sums to 1 over all values of parameter θ [11]. Thus, since $P(D)$ is constant, we can write Bayes' theorem in

one of the most useful form in Bayesian update and inference as follows:

$$\begin{aligned} P(\theta \mid D) &\propto P(D \mid \theta) \times P(\theta) \\ Posterior &\propto Likelihood \times Prior \end{aligned} \tag{2}$$

In Bayesian analysis, the results of the experiment could be used to update the belief about the parameter θ. In simple cases, we can compute the posterior distribution for the parameter θ by multiplying the prior distribution and the likelihood function as shown in equation 2. However, it is convenient mathematically for the prior and the likelihood to be conjugate. A prior distribution is a *conjugate prior* for the likelihood function if the posterior distribution belongs to the same distribution as the prior [21]. For example, the beta distribution is a conjugate prior for the binomial distribution (as a likelihood function) because the posterior distribution obtained by multiplying the prior and the likelihood belongs to the same distribution as the prior (thus, both the prior and the posterior have beta distributions).

Let's consider the effect of different priors on the posterior distribution. A completely uninformative prior is the beta distribution with parameters $\alpha = 1$ and $\beta = 1$. The posterior distribution, in this case, is equivalent to the likelihood function since we have a completely uninformative prior. More informative priors will have a greater influence on the posterior distribution for a given sample size. On the other hand, larger sample sizes will give the likelihood function more influence on the posterior distribution for a given prior distribution. In practice, this means that we can obtain a precise estimate of the posterior distribution using smaller sample sizes when we use more informative priors. Similarly, we may need larger sample sizes when we use a weak or uninformative prior.

$$\begin{aligned} P(\theta \mid D) &\propto Beta(\alpha, \beta) \times Binomial(n, \theta) \\ P(\theta \mid D) &\propto Beta(y + \alpha, n - y + \beta) \end{aligned} \tag{3}$$

Equation 3 is the formula that we are going to use in this paper for prior to posterior update. Starting with a prior distribution $Beta(\alpha, \beta)$, we add the count of successes,y, and the count of failures, $n - y$, from the dataset D (where n is total number of entries in D) to α and β, respectively [21]. Thus, $Beta(y+\alpha, n-y+\beta)$ is the posterior distribution.

2.3 Link Strengths in Bayesian Networks

The concept of link strength in Bayesian networks was introduced first by Boerlage in 1992 [4]. In his thesis, Boerlage introduced the concepts of both connection strength and link strength in a binary Bayesian network model. *Connection strength* for any two variables A and B in a Bayesian network model B_1 is defined as measuring the strength between these two variables by testing all possible paths between them in B_1, whereas *link strength* is defined as measuring the strength these two random variables taking into account only the direct edge $A - B$ [4]. Methods for link strengths measurements are not studied sufficiently. Imme Ebert-Uphoff in her 2009 paper [6] presented a tutorial on how to measure connection strengths and link strengths in discrete Bayesian networks. Ebert-Uphoff concluded that there is limited literature on link strengths, and there is more

need to apply and use link strengths measures in structure learning and other purposes [6]. However, to the authors' best knowledge, there are no more recent publications that address link strengths measurements in discrete Bayesian networks. In this paper, we define a novel and not computationally expensive link strengths measure in discrete Bayesian networks.

In this paper, we propose a link strengths measure (denoted by L_S) for discrete Bayesian networks. We then use L_S to determine the weakest edge and the most believable edge in a given causal model. We further study the robustness of the PC algorithm against data poisoning attacks that aim to remove the weakest edge and insert the most believable yet incorrect edge.

3 Measuring Link Strengths from Data in Discrete Bayesian Networks

In this paper, we introduce a novel link strengths measure between two random variables in a discrete Bayesian network model. It is essential to not only study the existence of a link in a causal model but also define a reliable link strengths measure that is useful in Bayesian reasoning [4, 6]. The new defined link strengths measure assigns a number to every link in a Bayesian network model. This number represents the lowest confidence of all possible combinations of assignments of posterior distributions. The defined link strengths measure will be used to rank edges from the most to the least believable edge, rank edges from the weakest to the strongest edge, and justify a plausible process in any causal model. Our novel approach is as follows:

Definition 1. *Link Strengths Measure L_S is defined as*

$$L\_S(Variable_1 \rightarrow Variable_2) = \min_{y \in Y}(pdf(\frac{y+\alpha}{\alpha+n+\beta})) \tag{4}$$

where $Y = \{n_{11}, n_{12}, \cdots, n_{1j}, n_{21}, n_{22}, \cdots, n_{2j}, \cdots, n_{i1}, n_{i2}, \cdots, n_{ij}\}$, pdf is the probability density function, and $\frac{y+\alpha}{\alpha+n+\beta}$ is the mean of the posterior distribution.

Explanation: Given a discrete dataset DB_1 and a Bayesian network structure B_1 learned by the PC algorithm using DB_1, for every link $Variable_1 \rightarrow Variable_2$ in B_1, build a contingency table [13] for the two discrete variables $Variable_1$ and $Variable_2$ with i and j states, respectively (as shown in table 1). To measure the strength of links of a causal model, we perform the following two steps:

(1) We compute the posterior distributions for each link $Variable_1 \rightarrow Variable_2$ as follows: $P(Variable_2 \mid Variable_1) = Beta(y+\alpha, n-y+\beta)$ where $variable_2 \mid variable_1$ is all possible combinations of discrete states of $Variable_2$ and $Variable_1$, and then

(2) We use our link strengths measure as presented in equation 4. Note that $\frac{y+\alpha}{\alpha+n+\beta}$ in equation 4 is obtained by simply substituting α with $y+\alpha$ and β with $n-y+\beta$ in $\frac{\alpha}{\alpha+\beta}$.

Interpretation: For any two random variables in a causal model ($variable_1$ with i states and $variable_2$ with j states), there are $i \times j$ combinations of assignments of posterior distributions. For every posterior distribution, we have a prior distribution that is a conjugate prior for the likelihood function. For instance, a posterior distribution in the form $Beta(y+\alpha, n-y+\beta)$ has a Beta-distributed prior, $Beta(\alpha, \beta)$, which is a conjugate prior for the likelihood function, $Binomial(n, \theta)$. Considering all $i \times j$ posterior distributions for the two random $variable_1$ and $variable_2$, we can measure the uncertainty of that link by measuring how peaked the posterior distributions (Beta distributions in our experiments) are; thus, we can identify the link strength based on the uncertainty level. The more peaked the posterior distribution is, the more certainty we have about the posterior distribution probability. In other words, the peak of a beta distribution,$Beta(\alpha', \beta')$, is reached at its mean, $\frac{\alpha'}{\alpha'+\beta'}$. Thus, the peak of the posterior distribution is reached at $\frac{y-\alpha}{n-y+\beta}$. In the defined link strengths measure, we define the link strength for any link between two random variables in a causal model as the value of the smallest peak. This point is the point at which the model has seen the fewest number of cases; thus, it is the most critical point through which this link can be manipulated.

	Variable₂			
Variable₁	**State₁**	$\cdots$	**State$_j$**	**Observed Row Total**
State₁	$[n_{11}], (e_{11}), < ts_{11} >$	$\cdots$	$[n_{1j}], (e_{1j}), < ts_{1j} >$	$\sum_{t=1}^{j} n_{1t}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
State$_i$	$[n_{i1}], (e_{i1}), < ts_{i1} >$	$\cdots$	$[n_{ij}], (e_{ij}), < ts_{ij} >$	$\sum_{t=1}^{j} n_{it}$
Observed Column Total	$\sum_{t=1}^{i} n_{t1}$	$\cdots$	$\sum_{t=1}^{i} n_{tj}$	n (Observed Grand Total)

Table 1: A contingency table for two discrete variables $Variable_1$ and $Variable_2$ with i and j states, respectively. The contingency table is structured as follows: $[n_{ij}]$ is the cell's observed counts obtained from dataset DB_1, (e_{ij}) is the cell's expected counts, calculated as follows: (*Observed Row Total* $\times$ *Observed Column Total*) $\div$ (*Observed Grand Total (denoted as n)*), and $< ts_{ij} >$ is the cell's chi-square test statistic, calculated as follows: $(n_{ij} - e_{ij})^2 \div e_{ij}$.

Practical usages: We use this measure to identify weak edges (i.e., low values of L_S). These edges are the easiest to remove from a given causal model. We also use the L_S value to identify location for new edges to be added. We claim that the highest L_S value, the most believable the new edge is. A practical usage of the proposed link strengths measure is that it can be used to evaluate the robustness of the PC algorithm against data poisoning attacks.

4 Data Poisoning Attacks against the PC Algorithm

In the process of using data to learn the structure of a Bayesian network model, the PC algorithm assesses conditional independence statements linking variables. The χ^2 statistical test is conducted on the given dataset to outline the statistical independence statement

set for the learned causal model [18]. As a result of delineating how the PC algorithm works, adversarial attackers may exploit this knowing by contaminating the input dataset via weak edges removal or insertion of believable, yet incorrect links.

In this paper, we use our link strengths measure to investigate the robustness of the PC algorithm against two types of data poisoning attacks as follows: 1) Data poisoning attacks based on removing the weakest edge and 2) Data poisoning attacks based on inserting the most believable yet incorrect edge.

Due to space limitation, we only present selected algorithms in this work. A complete set of algorithms and further details can be accessed in our technical report [1].

4.1 Data Poisoning Attacks based on Removing the Weakest Edge

As discussed, it is feasible to use link strengths measure to identify and rank causal model edges from weakest to strongest, which means that adversarial opponents may seize the opportunity to poison the learning dataset, the objective being to effectively remove weak edges.

We have developed *Algorithm 4* to check the resilience of the PC algorithm against attacks that target weak edges. Our algorithm calculates the strength of each link in a Bayesian model and then rank the edges from the weakest to the strongest edge. It then checks the robustness of the PC algorithm against the feasibility of deleting the weakest edge. Our empirical results are presented in section 5.

Algorithm 4: Removing a Weak Edge Procedure

Input : Dataset DB_1 ▷ Original dataset with n cases
Output: Contaminated dataset DB_2 or a failure message

Procedure `Removing a Weak Edge(`DB_1`)`
  Use the PC algorithm for learning the structure of Bayesian network model B_1 from dataset DB_1 (using the default significance level at 0.05 [12])
  Use L_S to rank the edges of B_1 from the weakest to the strongest
  Let $A - C$ be the weakest edge to be deleted from B_1
  Test the feasibility of deleting the edge $A - C$ from B_1 using *Algorithm 3*
  if *Algorithm 3 returns* DB_2 **then**
    Return DB_2
  else
    Return msg "*Algorithm 3* failed to delete the link $A - C$ within a feasible number of cases"
  end
end

4.2 Data Poisoning Attacks based on Adding the Most Believable yet Incorrect Edge

We demonstrate that the use of link strengths measure can be successfully applied to a causal model to accurately identify and rank the edges from most to least in terms of

believability. Therefore, adversaries can skillfully use data poisoning attacks to generate input dataset to the Bayesian network model so that integrating the incorrect, yet plausible edges is viable.

Algorithm 6: Adding the Most Believable yet Incorrect Edge Procedure

Input : Dataset DB_1 ▷ Original dataset with n cases
Output: Contaminated dataset DB_2 or a failure message

Procedure `Adding the Most Believable yet Incorrect Edge(`DB_1`)`
  Use the PC algorithm for learning the structure of Bayesian network model B_1 from dataset DB_1 (using the default significance level at 0.05 [12])
  Choose a set of edge Q that could be added to B_1
  Use L_S to rank the set of edges Q from the most to the least believable edge
  Let $A - C$ be the most believable edge to be added to B_1
  if *$A - C$ lies in a a serial or diverging triple $A - B - C$* **then**
    Use *Algorithm 1* to check the feasibility of adding the link $A - C$
    if *Algorithm 1 returns* DB_2 **then**
      Return DB_2
    else
      Return msg "*Algorithm 1* failed to introduce the link $A - C$"
    end
  else if *$A - C$ lies in a converging triple $A \rightarrow B \leftarrow C$* **then**
    Use *Algorithm 2* to check the feasibility of adding the link $A - C$
    if *Algorithm 2 returns* DB_2 **then**
      Return DB_2
    else
      Return msg "*Algorithm 2* failed to introduce the link $A - C$"
    end
  else
    Use *Algorithm 5* to check the feasibility of adding the link $A - C$
    if *Algorithm 5 returns* DB_2 **then**
      Return DB_2
    else
      Return msg "*Algorithm 5* failed to introduce the link $A - C$"
    end
  end
end

We have developed *Algorithm 6* to check the robustness of the PC algorithm against this attack. The algorithm starts by learning the structure of the Bayesian network model and then uses the defined link strengths measure to rank a given set of edges that could be added to the learned model from the most to the least believable edge. Our algorithm then checks the robustness of the PC algorithm against the feasibility of adding the most believable edge. Our empirical results are presented in section 5.

5 Empirical Results

In this section, we present the results of using our link strengths measure on the Chest Clinic Network [10] and then demonstrate some of its practical usages. We implemented the Chest Clinic Network (shown in Figure 1) using *Hugin™ Research 8.1*. Then we simulated dataset of $10,000$ cases for our experiments by using *Hugin™ case generator* [12, 19]. We call this dataset as DB_1. Using the PC algorithm on dataset DB_1 with 0.05 significance setting [12], the resulting structure is given in Figure 2. While the two networks belong to different Markov equivalence classes, we will use the network of Figure 2 as the starting point of our experiments.

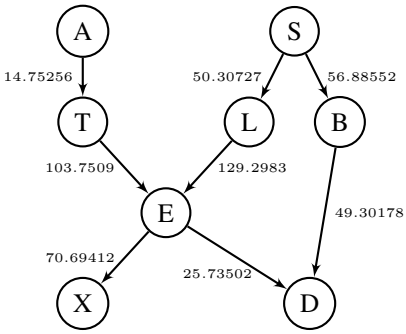


Figure 1: Chest Clinic Network and the result of link strengths (L_S)

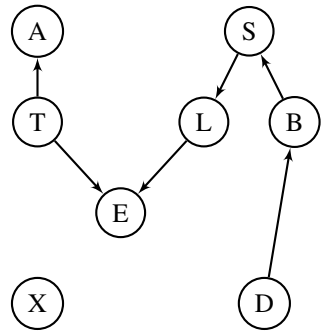


Figure 2: B_1, the result of feeding DB_1 to the PC algorithm with significance level at 0.05

We computed the link strengths using our approach (shown in Figure 1). We evaluated the effectiveness of data poisoning attacks against the PC algorithm (presented in section 4) to poison the Chest Clinic Network dataset DB_1 as follows: First, we validate the effectiveness of data poisoning attacks based on removing the weakest edge described in *Algorithm 4* to contaminate DB_1. Second, we check the resilience of the PC algorithm against the feasibility of data poisoning attacks based on adding most believable yet incorrect edge described in *Algorithm 6* to poison DB_1.

We present our results of deleting the weakest edge from B_1 in Figure 3. We observe that *Algorithm 4* succeeded to determine the weakest edge, $A - T$, and delete it by modifying only 3 cases in our dataset DB_1. Our results of adding the most believable edge to B_1 are presented in Figure 4. We observe that *Algorithm 6* succeeded to fool the PC algorithm and introduce the most believable edge, $B - L$, from the set of edges Q (in our experiment, we let $Q = \{A - S, T - S, D - S, L - B, L - T\}$) by inserting only 13 corrupt cases to our dataset DB_1.

We observed that when removing an edge from a causal model, the choice of corrupt data items has an impact on the efficiency of the attack. That is, transferring data items from the cell with the highest test statistics value to the cell with the lowest test statistics value in a contingency table of two random variables will accelerate the process of removing the link between them. We also observe that when introducing a new malicious link

between two random variables, a cell with a higher test statistics value $< ts_{ij} >$ in the contingency table of these two random variables requires fewer corrupt cases than a cell with a lower test statistics value. Overall, we showed that the PC algorithm is vulnerable to data poisoning attacks based removing the weakest edge and adding the most believable yet incorrect edge.

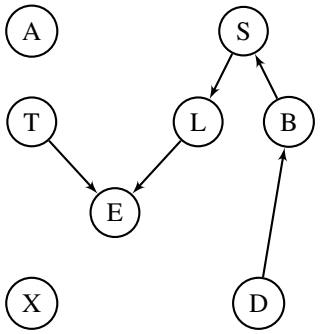


Figure 3: The result of removing the weakest link in B_1, $A \rightarrow T$

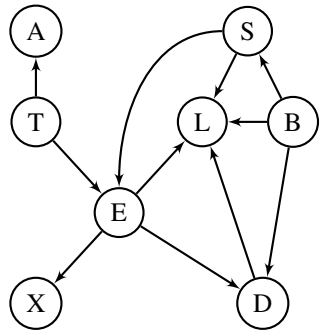


Figure 4: The result of adding the most believable link to B_1, $B \rightarrow L$.

6 Conclusion and Future Work

In this paper, we demonstrated the vulnerability of the PC structure learning algorithm. We have developed a theoretical framework to classify data poisoning attacks against the PC algorithm. We also performed experimental studies using the widely used Chest Clinic dataset. Our findings indicate that the PC algorithm is highly sensitive to data poisoning attacks. We also demonstrated that attackers could corrupt the learning outcome in a way that the PC algorithm will learn the desired structure.

Our novel link strength measure plays a crucial role in identifying vulnerable network structure and the ease of corrupting the Bayesian model. We believe that using this measure will guide defensive measurements. Our ongoing work includes the development of methods that will reduce the risk of unauthorized compromise against the PC algorithm via data poisoning.

References

[1] E. Alsuwat, M. Valtorta, and C. Farkas. Bayesian structure learning attacks. Technical report, University of South Carolina, SC, USA, 2018.

[2] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, Nov 2010.

[3] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1467–1474. Omnipress, 2012.

[4] B. Boerlage. *Link strength in bayesian networks*. PhD thesis, University of British Columbia, 1992.

[5] C. Burkard and B. Lagesse. Analysis of causative attacks against svms learning from data streams. In *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*, pages 31–36. ACM, 2017.

[6] I. Ebert-Uphoff. Tutorial on how to measure link strengths in discrete bayesian networks. Technical report, Georgia Institute of Technology, 2009.

[7] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58. ACM, 2011.

[8] A. Hugin Expert. S, 2008. *Hugin Researcher API 7.0 (www. hugin. com)*.

[9] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017.

[10] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

[11] S. M. Lynch. *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science & Business Media, 2007.

[12] A. L. Madsen, F. Jensen, U. B. Kjaerulff, and M. Lang. The hugin tool for probabilistic graphical models. *International Journal on Artificial Intelligence Tools*, 14(03):507–543, 2005.

[13] M. L. McHugh. The chi-square test of independence. *Biochemia medica: Biochemia medica*, 23(2):143–149, 2013.

[14] S. Mei and X. Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, pages 2871–2877, 2015.

[15] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 27–38. ACM, 2017.

[16] R. E. Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[17] A. Newell, R. Potharaju, L. Xiang, and C. Nita-Rotaru. On the practicality of integrity attacks on document-level sentiment analysis. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, pages 83–93. ACM, 2014.

[18] T. D. Nielsen and F. V. Jensen. *Bayesian networks and decision graphs*. Springer Science & Business Media, 2009.

[19] K. G. Olesen, S. L. Lauritzen, and F. V. Jensen. ahugin: A system creating adaptive causal probabilistic networks. In *Uncertainty in Artificial Intelligence, 1992*, pages 223–229. Elsevier, 1992.

[20] N. Papernot, P. McDaniel, and I. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[21] H. Raiffa and R. Schlaifer. *Applied statistical decision theory*. Div. of Research, Graduate School of Business Administration, Harvard Univ., 1961.

[22] M. Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3):1–22, 2010.

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

[24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[25] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli. Support vector machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015.

[26] H. Xiao, H. Xiao, and C. Eckert. Adversarial label flips attack on support vector machines. In *ECAI*, pages 870–875, 2012.

[27] C. Yang, Q. Wu, H. Li, and Y. Chen. Generative poisoning attack method against neural networks. *arXiv preprint arXiv:1703.01340*, 2017.

An Efficient Way to Compose Distributions from Exponential Families

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

Abstract

In previous works we attempted to compose multivariate densities of continuous random variables. The paper [5] showed an implementation of Iterative Proportional Fitting Procedure numerically approximating the multivariate density from low-dimensional ones. In [2] we defined an operation of composition for general continuous densities with advantageous properties in case of certain copula classes. The paper [4] further analyzed properties of composition in continuous densities and sketched a basic application for densities from exponential families. This application was illustrated on an non-trivial example in [3].

The exponential families are closed with respect to the operation of composition, i.e. the result of composition remains in the exponential family. Though up to this moment only rather toy applications were performed, still it took a non-trivial effort to perform algebraic manipulations with multivariate densities expressed in a "natural form" of exponential families. Therefore, it appears to be advantageous to employ some computer algebra system capable of symbolic manipulations with matrices necessary for the representation of multivariate distributions.

Keywords: Operator of composition, continuous variable, exponential family, computer algebra system.

1 Introduction

Modern equipment, electronic sensors and automated approaches of measurement provide an enormous amounts of data which is almost in every case multivariate and usually shows high dimensionality. This is the setting in which the curse of dimensionality (see, e.g., [1, 6]) appears to be serious issue. One particular facet of this is

a impossibility, or at least enormous inaccuracy of estimation of high-dimensional multivariate distributions from the data. A possible way of dealing with these problems is the employment of some factorization of high-dimensional distributions and performance of local computations with low-dimensional marginals only. Well-known are approaches of Bayesian Networks now successfully implemented in several commercial computational environments (Hugin, Bayesia, etc.). An algebraic alternative is represented by compositional models (see Jiroušek[11]).

As we already hinted, it appears to be impossible to estimate multivariate densities of higher dimensions directly from data. But still it is possible to perform some analysis of dependence structure among analyzed variables and to estimate the low-dimensional discrete distributions or continuous densities as basic building blocks. These building blocks overlap in a way, i.e. the considered marginals have some common variables and thus the building blocks can be seen as hypergraph edges, where hypergraph vertices are the particular variables.

Since we present a practical application of scheme sketched in previous publications ([4] and [3]), the methodology is connecting several different fields from theory of compositional models developed under the framework of classical probability theory (see Jiroušek[11, 4]), theoretical description of exponential families and its useful properties [13], employment of algorithms from theory of maximum likelihood estimation of multivariate normal distribution [7] with an implementation in R package *mvnmle* [9] and a Python based computer algebra system *SymPy* with implementation [10] under statistical computational environment of R [15].

The presented paper shows all necessary theoretical prerequisites concerning composition in exponential families together with the way how to implement compositional models in a rather user friendly way leaving the boring computations on a computer algebra system.

2 Compositional Models & Exponential Families

Within the presented paper we consider a finite index set $N = \{1, \dots, n\}$ together with a set of random variables $\{X_i\}_{i \in N}$ with values, or vectors of values, denoted by the corresponding lowercase letters. The domain of variables will be denoted by the corresponding bold uppercase letter $\mathbf{X}_i$. In general, variables with a finite or countable set of possible *states* are called *discrete*; other variables are called *continuous*. In this paper, we will focus on the later case.

The probability density functions of continuous random variables will be denoted by lowercase letters of the Latin alphabet ($f, g, h, \dots$), e.g., the abbreviated notation $f(x_K)$ denotes a multidimensional density of variables having indices from set $K \subseteq N$. For a probability density function $f(x_K)$ and any set of variable indices $L \subset K$, a marginal probability density $f(x_L)$ of $f(x_K)$ can be computed for each x_L as follows

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

where obviously the integration runs over the domains of all variables in $K \setminus L$. We will also employ an equivalent way to denote the marginal $f(x_L)$, namely $f^{\downarrow\{L\}}$ which was introduced by Glenn Shafer (see, e.g., [17]).

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

$$f(x_L \,|\, x_M = \mathbf{x_M}) f(x_M = \mathbf{x_M}) = f(x_L, x_M = \mathbf{x_M}).$$

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

2.1 Composition of Continuous Densities

Let us have two probability density functions $f(x_K)$ and $g(x_K)$ with the same set of variables X_K. Then f is said to be *absolutely continuous with respect to* g, or *dominated by* g (denoted by $f \ll g$) if for each $x_K \in X_K$ it holds

$$(g(x_K) = 0 \Rightarrow f(x_K) = 0)\,.$$

Consider two sets of continuous variables X_L and X_M, a probability density $f(x_L)$, and a probability density $g(x_M)$ such that $f(x_{L\cap M}) \ll g(x_{L\cap M})$. The *right composition* is given by

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

For details and important properties of composition in continuous case, please, refer to [4].

2.2 Exponential Families

The possibility to define the operation of composition for densities of distributions from *exponential families* was studied in [4]. The exponential family is an interesting set of probability distributions that can be expressed in a certain form, e.g., see [12].

Now, let us recall the most important notions and properties introduced in the context of compositional model in [4]. *Density* $f(x_L)$ *belongs to the exponential family* if it can be expressed in the form

$$f(x_L; \theta_L) = h(x_L) e^{\eta_L(\theta_L)\cdot T(x_L) - A(\eta_L)}$$

where θ_L is a (real) vector of parameters and $h(x_L)$, $T(x_L)$, $\eta_L(\theta_L)$ and $A(\eta_L)$ are vector functions.

The function $\eta_L(\theta_L)$ is a *natural parameter* (or exponential parameter), $T(x_L)$ is a *sufficient statistic*, $A(\eta_L)$ is a log-partition function and $h(x_L)$ is a non-negative base measure. Obviously, the product of $\eta_L(\theta_L)$ and $T(x_L)$ vector functions is

a scalar product. Examples of the most important members of the exponential family, such as Gaussian, binomial, multinomial, Gamma and Beta distributions can be found, e.g., in [13].

It can be shown that exponential family is closed with regard to several important operations, particularly product, marginalization and conditioning, see, e.g., Lemmata 6 and 8 in [12].

If both operands belong to the exponential family, the result of operation of composition is defined and can be expressed in the above form and thus also belongs to the exponential family. I.e. for two densities $f(x_L)$ and $g(x_M)$ belonging to an exponential family, i.e. such that $f(x_L) = h_L(x_L)e^{\eta_L \cdot T_L(x_L) - A_L(\eta_L)}$ and $g(x_M) = h_M(x_M)e^{\eta_M \cdot T_M(x_M) - A_M(\eta_M)}$ the *composition also belongs to the exponential family.*

Let us look at this property in more detail:For disjoint L and M we get the product of both densities, which obviously also belongs to the exponential family.

If the other possibility realizes, i.e. if $L \cap M \neq \emptyset$ then we can express

$$g(x_M) = h_M(x_M)e^{\eta_{L\cap M} \cdot T_{L\cap M}(x_{L\cap M}) + \eta_{M\setminus L} \cdot T_{M\setminus L}(x_{M\setminus L}) - A_M(\eta_{L\cap M}, \eta_{M\setminus L})}.$$

According to [12] the conditional distribution

$$g(x_{M\setminus L} \,|\, x_{L\cap M} = \mathbf{x_{L\cap M}}) = h_{\mathbf{L\cap M}}e^{\eta_{M\setminus L} \cdot T_{M\setminus L}(x_{M\setminus L}) - A_{\mathbf{L\cap M}}(\eta_{M\setminus L})}$$

where $h_{\mathbf{L\cap M}}$ and $A_{\mathbf{L\cap M}}$ are dependent on the values of conditioning variables. It is now apparent that the product of $f(x_L)$ and $g(x_{M\setminus L} \,|\, x_{L\cap M})$ again belongs to the exponential family since it can be written in the corresponding form, i.e.

$$(f \triangleright g)(x_{L\cup M}) = h_L h_{\mathbf{L\cap M}} e^{\eta_L \cdot T_L(x_L) + \eta_{M\setminus L} \cdot T_{M\setminus L}(x_{M\setminus L}) - A_L(\eta_L) - A_{\mathbf{L\cap M}}(\eta_{M\setminus L})}.$$

2.3 Multivariate Normal Distribution

The non-degenerate multivariate normal distribution has a symmetric and positive definite covariance matrix $\mathbf{\Sigma}$. In such case, the multivariate normal distribution $f(x_L)$ with vector of means μ_L and covariance matrix $\mathbf{\Sigma}_L$ has a density given by formula

$$f(x_L; \mu_L, \mathbf{\Sigma}_L) = \frac{1}{\sqrt{(2\pi)^\ell |\mathbf{\Sigma}_L|}} \exp\left(-\frac{1}{2}(x_L - \mu_L)^{\mathrm{T}} \mathbf{\Sigma}_L^{-1} (x_L - \mu_L)\right)$$

where ℓ is a dimension (length) of x_L vector, symbol $^{\mathrm{T}}$ stands for a vector transpose, $|\mathbf{\Sigma}_L|$ is determinant of covariance matrix and $\mathbf{\Sigma}_L^{-1}$ is an inverse of covariance matrix.

Thus, multivariate density $f(x_L; \mu_L, \mathbf{\Sigma}_L)$ has variables and functions according

to definition of exponential family given in the following way

$$\begin{aligned} x_L &= (x_1, \ldots, x_\ell)^{\mathrm{T}}, \\ \eta_L &= \begin{pmatrix} \boldsymbol{\Sigma}_L^{-1}\mu_L \\ -\frac{1}{2}\boldsymbol{\Sigma}_L^{-1} \end{pmatrix}, \\ T_L(x_L) &= \begin{pmatrix} x_L \\ x_L x_L^{\mathrm{T}} \end{pmatrix}, \\ A_L(\eta_L) &= \frac{1}{2}\mu_L^{\mathrm{T}}\boldsymbol{\Sigma}_L^{-1}\mu_L + \frac{1}{2}\log|\boldsymbol{\Sigma}_L|, \\ h_L(x_L) &= (2\pi)^{-\frac{\ell}{2}}. \end{aligned}$$

2.4 Conditional multivariate density

Let us have a multivariate density $g(x_M; \mu_M, \boldsymbol{\Sigma}_M)$ and let us divide index set M into two disjoint parts such that $A = L \cap M$ and $B = M \setminus L$. Thus, the m-dimensional vector x_M can be partitioned into two parts of dimensions m_A and m_B ($m_A + m_B = m$) in such a way that

$$x_M = \begin{pmatrix} x_A \\ x_B \end{pmatrix}$$

and similarly

$$\mu_M = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}.$$

The covariance matrix is partitioned into the corresponding blocks in the following way

$$\boldsymbol{\Sigma}_M = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix}$$

having sizes

$$\begin{pmatrix} m_A^2 & m_A m_B \\ m_A m_B & m_B^2 \end{pmatrix}.$$

Thus, having the multivariate density $g(x_M) \sim \mathcal{N}(\mu_M, \boldsymbol{\Sigma}_M)$ the *conditional multivariate density* $g(x_{M\setminus L} \,|\, x_{L\cap M} = \mathbf{a}) = g(x_B \,|\, x_A = \mathbf{a})$ is again a multivariate density distribution (see, e.g., [8]) and $g(x_B \,|\, x_A = \mathbf{a}) \sim \mathcal{N}(\overline{\mu}_B, \overline{\boldsymbol{\Sigma}}_B)$ where

$$\overline{\mu}_B = \mu_B + \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}(\mathbf{a} - \mu_A)$$

and

$$\overline{\boldsymbol{\Sigma}}_B = \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{AB}.$$

We can somewhat surprisingly see, that the known value $\mathbf{a}$ influences the mean of conditional density but not its covariance matrix. Let us note that the formula for $\overline{\boldsymbol{\Sigma}}_B$ is known as the Schur complement of $\boldsymbol{\Sigma}_{AA}$ in $\boldsymbol{\Sigma}_M$ and $\boldsymbol{\Sigma}_{AA}^{-1}$ is a generalized inverse (see again [8]).

2.5 Product of Multivariate Densities

Similarly, the product of two multivariate normal densities is again multivariate normal distribution (must be then renormalized). For two multivariate densities $f(x_L) \sim \mathcal{N}(\mu_L, \mathbf{\Sigma}_L)$ and $g(x_M) \sim \mathcal{N}(\mu_M, \mathbf{\Sigma}_M)$ we get

$$f(x_L)g(x_M) \sim \mathcal{N}(\overline{\mu}, \overline{\mathbf{\Sigma}})$$

where

$$\overline{\mu} = \overline{\mathbf{\Sigma}} \left(\mathbf{\Sigma}_L^{-1} \mu_L + \mathbf{\Sigma}_M^{-1} \mu_M \right)$$

and

$$\overline{\mathbf{\Sigma}} = \left(\mathbf{\Sigma}_L^{-1} + \mathbf{\Sigma}_M^{-1} \right)^{-1}.$$

The normalizing constant is (see [16]) equal to

$$(2\pi)^{-\frac{\ell+m}{2}} \left| \mathbf{\Sigma}_L + \mathbf{\Sigma}_M \right|^{\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\mu_L - \mu_M \right)^{\mathrm{T}} \left(\mathbf{\Sigma}_L + \mathbf{\Sigma}_M \right)^{-1} \left(\mu_L - \mu_M \right) \right).$$

3 Partially Symbolic Manipulation with Compositional Models

As the kind reader already guessed from the formulas in previous section, the general case of composition in exponential families involves several matrix operations with partially numeric and partially symbolic manipulation. Obviously, it is advantageous to perform all computations in an (semi)automated way. We performed all implementations of compositional models in R software [15] which is very advantageous for its vector and matrix operations together with abundance of statistic and probabilistic methods available. Therefore, we decided to employ a Python based computer algebra system SymPy with its R interface rSymPy [10].

First of all, let us describe a simple data set which will be used in the following application of above described theory. It concerns the levels of 5 characteristics measured in the folicular fluid of 22 pregnant cows. The five variables (pH, pCO_2, pO_2, HCO_3, BE(B)) appear to have Gaussian distribution (first three variables on the 5 percent significance level using Shapiro-Wilk test of normality, two other variables on 1 per thousand significance level which is cause in both cases by a pair outliers). Thus, it appears to be a bad idea to approximate a five-dimensional multivariate Gaussian density with 30 continuous parameters from 110 measurements of 5 variables and it seems to be a good idea to estimate from data several low-dimensional (two- or three-dimensional) distributions and to compose them. (Three-dimensional distribution has 12 parameters.)

In this paper, we will not focus on the choice of the most suitable marginals as building stones. We will choose them in a rather rough and intuitive way based on the Pearson correlation matrix of the five variables (see Table 1). The proper way is to use some principles of probabilistic structure learning approaches (see, e.g., Zhou [18]).

Table 1: (Pearson) correlation matrix of five analyzed variables.

	pH	pCO_2	pO_2	HCO_3	BE(B)
pH	1.000	-0.549	0.529	0.647	0.704
pCO_2	-0.549	1.000	-0.509	0.280	0.206
pO_2	0.529	-0.509	1.000	0.141	0.184
HCO_3	0.647	0.280	0.141	1.000	0.997
BE(B)	0.704	0.206	0.184	0.997	1.000

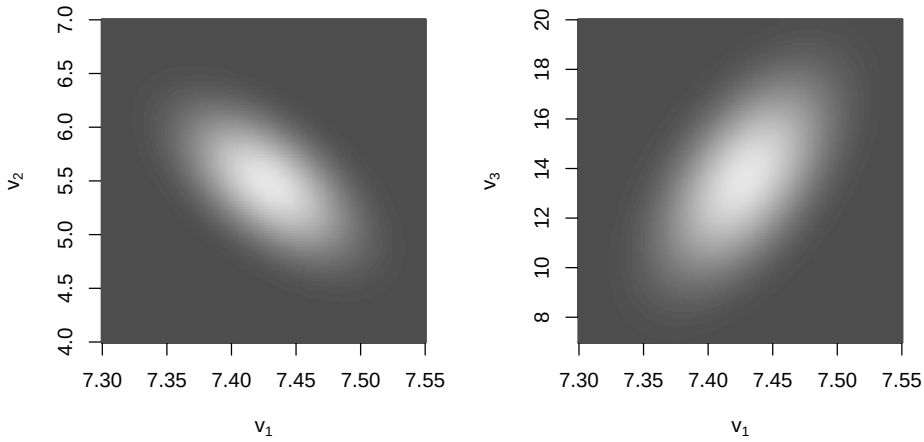


Figure 1: Two-dimensional Gaussian densities of variables of the first composed marginal v_1 and v_2 (left part of figure) and the second composed marginal v_1 and v_3 (right part of figure).

In the following text the five parameters will be denoted by $v_1, \ldots, v_5$ variable symbols. Rather loosely based on the strengthes of the Pearson correlations of moderate and strong linear dependencies we decided to choose three marginals $f_1(v_1, v_2)$, $f_2(v_1, v_3)$ and $f_3(v_1, v_4, v_5)$. For these marginals we estimate their multivariate Gaussian densities from data using maximum likelihood estimates specified in terms of the inverse of the Cholesky factor of the variance-covariance matrix (see [14]) and implemented in R *mvnmle* package [9]. Two estimated two-dimensional densities are depicted in Figure 1.

As the first step, we read our sample data set into `data` data frame and load the above mentioned libraries of *mvnmle* and *rSymPy*. Then we set the simplified variable names, define the list of marginals to be composed (using a list `edges`) and estimate parameters of multivariate distributions using `mlest` function (see bellow).

```
v <- 1:ncol(data)
names <- paste("v",v,sep="")
# names: "pH"   "pCO2" "pO2"   "HCO3" "BE.B"
edges <- list(c(1,2), c(1,3), c(1,4,5))

# MLE for multivariate normal distributions
mu <- NULL; sig <- NULL
for (e in 1:length(edges)) {
    est <- mlest(data[,edges[[e]]])
    emu <- est$muhat
    names(emu) <- names[edges[[e]]]
    mu <- c(mu, list(emu))
    esig <- est$sigmahat
    rownames(esig) <- names[edges[[e]]]
    colnames(esig) <- names[edges[[e]]]
    sig <- c(sig, list(esig))
}
```

The estimated parameters are stored in the following numbered vectors and matrices `mu*` and `sig*`.

```
> cat(sympy("mu1"))    > cat(sympy("sig1"),"\n")
[7.42868181808]        [ 0.0012049443106, -0.0069107030493]
[5.47636363472]        [-0.0069107030493,  0.1314049638269]
> cat(sympy("mu2"))    > cat(sympy("sig2"),"\n")
[7.42868181927]        [ 0.001204944283,   0.0393622676519]
[13.6313637645]        [ 0.039362267652,   4.5957484887845]
> cat(sympy("mu3"))    > cat(sympy("sig3"),"\n")
[7.42873958083]        [ 0.001204867353, 0.045509151565, 0.053034419262]
[26.7414588678]        [ 0.045509151565, 4.216007981078, 4.489984753482]
[2.75089979417]        [ 0.053034419262, 4.489984753481, 4.812928862933]
```

For all three distributions (hypergraph edges) we compute η_i stored in `e*`, T_i stored in `T*`, A_i stored in `A*` and h_i stored in `h*`. Corresponding densities f_i are computed and stored in `f*`.

```
for (e in 1:length(edges)) {
    sympy(paste("ea=(sig",e,")**(-1)*mu",e,"",sep=""))
    sympy(paste("eb=-((sig",e,")**(-1))/2",sep=""))
    sympy(paste("e",e,"=(ea.T).col_join(eb)",sep=""))

    sympy(paste("T",e,"=(x",e,".T).col_join(x",e,"*x",e,".T)",sep=""))
    sympy(paste("A",e,"=(mu",e,".T*((sig",e,")**(-1)*mu",e,")/2).det()+
          log(sig",e,".det())/2",sep=""))
    sympy(paste("l=1*x",e,".shape[0]",sep=""))
    sympy(paste("h",e,"=1/sqrt((2*pi)**l)",sep=""))

    sympy(paste("f",e,"=h",e,"*exp(e",e,".dot(T",e,")-A",e,")",sep=""))
}
```

Thus, we defined all three continuous densities $f_1, \ldots, f_3$ and all vector functions of an exponential family. We can continue in computation of conditional density $f_2(v_3 \mid v_1)$ using the formulae in subsection 2.4, i.e. the corresponding subvectors

μ_A and μ_B, submatrices Σ_{AA}, Σ_{AB}, Σ_{BA} and Σ_{BB} and functions defining the conditional density.

```
> sympy(paste("sigc=sigbb-sigba*sigaa**(-1)*sigab"))
[1] "[3.30988976740274]"
> sympy(paste("conda=Matrix([",paste("[",paste("v",A,sep=""),"]",
+ collapse=","),"])",sep=""))
[1] "[v1]"
> sympy(paste("muc=mub+sigba*sigaa**(-1)*(conda-mua)"))
[1] "[-229.043560103312 + 32.6672927676576*v1]"
> sympy("econd=(((sigc)**(-1)*muc).T).col_join(-(sigc**(-1))/2)")
     "[-69.1997547347452 + 9.86960142581775*v1]"
     "[                    -0.151062432629697]"
> cat(sympy("Tcond=(xcond.T).col_join(xcond*xcond.T)"),"\n")
     "[   v3]"
     "[v3**2]"
> sympy("Acond=(muc.T*((sigc)**(-1)*muc)/2).det()+log(sigc.det())/2")
[1] "7924.8790914+log(3.30988977)/2-2260.5686474*v1+161.2065797*v1**2"
> sympy("hcond=1/sqrt((2*pi)**(xcond.shape[0]))")
[1] "2**(1/2)/(2*pi**(1/2))"
```

The conditional density $f_2(v_3 \,|\, v_1)$ is then composed from the corresponding vector functions defining it as a member of exponential family. We arrive at a three-dimensional distribution $f_1(v_1, v_2) \triangleright f_2(v_1, v_3)$ defined by following density (where numeric values were rounded to two decimal places).

```
13.07*2**(1/2)*exp(-43420.44+523.94*v2+11430.68*v1-69.20*v3+9.87*v1*v3
                   -62.50*v1*v2-0.15*v3**2-5.45*v2**2-755.38*v1**2)/pi**(3/2)
```

In a similar manner we can compose the result above with a conditional distribution $f_3(v_4, v_5 \,|\, v_1)$ computed again using the formulae in subsection 2.4 from $f_3(v_1, v_4, v_5)$. Now, the result of the second composition is a five-dimensional distribution $(f_1(v_1, v_2) \triangleright f_2(v_1, v_3)) \triangleright f_3(v_1, v_4, v_5)$ defined again by density (rounded to two decimal places).

```
94.25*2**(1/2)*exp(-1382025.57+523.94*v2+36863.95*v4+252860.96*v1
-69.20*v3-37118.13*v5+9.87*v1*v3+3326.13*v1*v5+517.50*v4*v5-62.50*v1*v2
-3297.34*v1*v4-257.89*v4**2-11686.21*v1**2-259.82*v5**2-0.15*v3**2
-5.45*v2**2)/pi**(5/2)
```

From the resulting multivariate density we can symbolically express marginals which were not estimated from the data. This can be particularly useful in cases when we obtain estimates of the marginal building blocks in the process of composition from different data sources.

The computation of marginals can be hardly computationally (or algorithmically) feasible. The package *rSymPy* did not succeeded in symbolic integration of marginal $(f_1(v_1, v_2) \triangleright f_2(v_1, v_3))(v_2, v_3)$ but we succeeded in computation of the corresponding two-dimensional density using online tool of Wolfram Alpha (see Figure 2). The approximately computed result appears to be again in the form of Gaussian density, i.e. it belongs to the exponential family as expected according to the above presented assertions. But it appears that computation of marginals is in general rather uneasy task which probably can be made feasible if the integration procedure take advantage of exponential family properties.

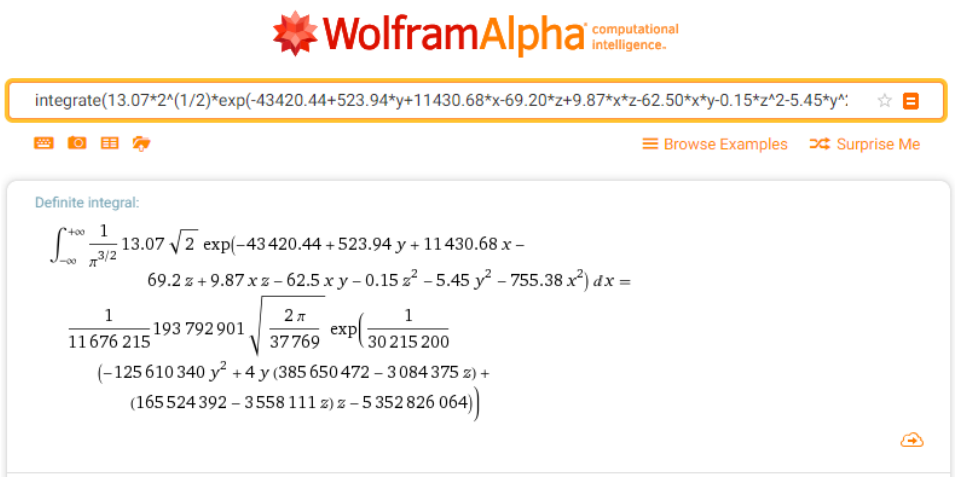


Figure 2: Two-dimensional Gaussian density computed as a marginal from the composition.

4 Conclusions and Possible Continuation

In this paper we presented an application of computer algebra system *rSymPy* able to perform operations with parameters of exponential families in order to (partially) symbolically perform the operation of composition using exponential families, namely low-dimensional Gaussian densities estimated from data.

Let us mention that the computer algebra system is capable of computation in arbitrary precision. Moreover, if the parameters of composed distributions are rational all operations are performed precisely and results is an exact expression (in a same way as in a toy example in [3]).

The possible future course of development contains the user friendly interface, setting of proper procedures for (sub)optimal choice of composed marginals and their estimation and elaboration of procedures leading to the efficient integration of marginals of compositional models.

Acknowledgement
The research was supported by Academy of Sciences of the Czech Republic and Ministry of Science and Technology (Taiwan) under project MOST-18-4 (Compositional models for data mining).

Author also wants to express his sincere gratitude for providing the data to associate professor Svatopluk Čech from Brno University of Veterinary and Pharmaceutical Sciences.

References

[1] R. Bellman. *Dynamic Programming*. Dover Books on Computer Science Series. Dover Publications, 2003.

[2] V. Bína. An Operator of Composition for the Multivariate Copula Construction. In *Proceedings of the 33rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt'13)*, volume 1636, pages 5–10, Canberra, Australia, dec 2014.

[3] V. Bína. Composition in exponential families: Case of multivariate normal distribution. In V. Novák, M. Inuiguchi, and M. Štěpnička, editors, *Proceedings of the 20th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty*, pages 9–14, Pardubice, 2017.

[4] V. Bína. Composition Operator for Densities of Continuous Variables. *International Journal of Approximate Reasoning*, 88:614–623, 2017.

[5] V. Bína and R. Jiroušek. A Short Note on Multivariate Dependence Modeling. *Kybernetika*, 49(3):420–432, 2013.

[6] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS Conference on Math Challenges of the 21st Century*, 2000.

[7] P. S. Dwyer. Some applications of matrix derivatives in multivariate analysis. *Journal of the American Statistical Association*, 62(318):607–625, 1967.

[8] M. L. Eaton. *Multivariate statistics: A vector space approach*. John Wiley and Sons, New York, 1983.

[9] K. Gross and with help from Douglas Bates. *mvnmle: ML estimation for multivariate normal data with missing values.*, 2012. R package version 0.1-11.

[10] G. Grothendieck and C. C. J. G. Bellosta. *rSymPy: R interface to SymPy computer algebra system*, 2012. R package version 0.2-1.1.

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

[12] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[13] F. Nielsen and V. Garcia. Statistical exponential families: A digest with flash cards. *CoRR*, abs/0911.4863, 2009.

[14] J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296, 1996.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[16] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[17] G. Shafer. *Probabilistic Expert Systems*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, USA, 1996.

[18] Y. Zhou. Structure learning of probabilistic graphical models: A comprehensive survey. *CoRR*, abs/1111.6925, 2011.

Behavior of L_1-based probabilistic correction applied to statistical matching with misclassification information

Giulia Buonumori **Andrea Capotorti**
Dip. Matematica e Informatica
University of Perugia - Italy
andrea.capotorti@unipg.it giulia.buonumori@studenti.unipg.it

Abstract

We illustrate the use of a recently proposed efficient procedure, based on L_1 distance minimization, for correcting inconsistent (i.e. incoherent) probability assessments for the so named statistical matching problem. Albeit the statistical matching problem is based on conditional probabilities estimates, inconsistencies can appear only among assessments given on the same conditioning values, hence a correction instance can be splitted in a finite set of unconditional correction instances where the L_1-based correction can efficiently operate. The statistical matching problem has been recently enriched with the possibility of a misclassification setting, breaking the aforementioned segmentation possibility. Anyhow the L_1-based procedure can be applied by a straightforward translation in a MIP problem, albeit the set of consistent solutions turns out to be not convex and hence potential disconnected solutions can appear.

1 Introduction

In recent contributions [1, 2] it has been proposed an efficient procedure for correcting inconsistent (i.e. incoherent) probability assessments based on L_1 distance minimization and encoded in mixed integer programming (MIP) problems. The procedure is particular apt to deal with assessments stemming from different sources of information, and the so named statistical matching problem is one of those cases (see e.g. [11]). Albeit the statistical matching problem is based on conditional probabilities estimates, always in [11] it has been proven that inconsistencies can appear only among assessments given on the same conditioning values, hence a correction instance can be splitted in a finite set of unconditional correction instances where the L_1-based correction can efficiently operate.

The problem has been recently enriched with the possibility of a misclassification setting [8], breaking the aforementioned segmentation possibility. If marginal assessments on the conditioning variable are taken for good, the only possible correction are the closest Fréchet-Hoeffding bounds for the misclassification probabilities. On the contrary, if also the marginal probabilities are allowed to be modified or the assessment is partial, the L_1-based procedure can be applied by a straightforward translation in a MIP problem, albeit the set of consistent solutions turns out to be not convex and hence potential disconnected solutions can appear. It is eventually notable that in the case the L_1-based correction would induce some marginal probability to be null, that could happen whenever the initial assessment would be based on rare or scarce observations, it will not be needed to proceed to further corrections on deeper zero layers (see [5]).

In the next sections we will briefly illustrate the general statistical matching (Sec.2), the merging and correction procedures for general unconditional probability assessments (Sec.3) and consequently their specific application to the statistical matching problem (Sec.4). Finally, in Sec.5 we will give a rough preliminary idea of the correction of incoherent evaluations when also a missclassification mechanism is assessed.

2 The statistical matching problem

As already stated, we propose to adopt a correction procedure applied to a merging operation for a specific practical problem named "statistical matching". Let us briefly recall what it means and which are its main peculiarities. A detailed description of such a problem can be found, e.g., in [9, 10].

Denote by $(\mathcal{X}_1, \mathcal{Y}_1), \ldots, (\mathcal{X}_{n_A}, \mathcal{Y}_{n_A})$ and by $(\mathcal{X}_{n_A+1}, \mathcal{Z}_{n_A+1}), \ldots, (\mathcal{X}_{n_A+n_B}, \mathcal{Z}_{n_A+n_B})$ two random samples, related to two sources A and B, of dimensions n_A and n_B. Samples observe three categorical variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ with modalities mx_i, $i \in I$, my_j, $j \in J$ and mz_k, $k \in K$, respectively. Hence in the sequel we will adopt the following notation for the possible observations:

$$X_i \equiv (\mathcal{X} = mx_i), i \in I, \quad Y_j \equiv (\mathcal{Y} = my_j), j \in J, \quad Z_k \equiv (\mathcal{Z} = mz_k), k \in K, \quad (1)$$

that will constitute our propositional variables (i.e. events).

Let S_s (with $s = 1, 2$) be the two, possibly different, sampling schemes. From them, relevant parameters, represented by (conditional) probabilities, can be estimated : from A the probability to observe Y_j conditional on X_i (for any $i \in I$)

$$\mathbf{y}_{j|i} = P_{\mathcal{Y}|(X_i)}(Y_j), \quad (2)$$

and analogously from B the probability to observe Z_k conditional on X_i (for any $i \in I$)

$$\mathbf{z}_{k|i} = P_{\mathcal{Z}|X_i}(Z_k). \quad (3)$$

Moreover, from A we can estimate the probability to observe X_i by following the first sampling scheme

$$\mathbf{x}_i^{S_1} = P_{\mathcal{X}}(X_i|S_1), \tag{4}$$

while from file B by following the second one

$$\mathbf{x}_i^{S_2} = P_{\mathcal{X}}(X_i|S_2), \tag{5}$$

and, by supposing that an observation can be obtained through one single sampling scheme S_s, with $s \in \{1, 2\}$ and probability $P(S_s)$, we get

$$\mathbf{x}_i = P_{\mathcal{X}}(x_i) = \mathbf{x}_i^{S_1} P(S_1) + \mathbf{x}_i^{S_2} P(S_2). \tag{6}$$

Under the assumption of a common sampling scheme, estimations are obtained through partial maximum likelihood method, and the result brings to the frequencies

$$\mathbf{y}_{j|i} = \frac{n_A^{ij}}{n_A^{i\cdot}} \quad , \quad \mathbf{z}_{k|i} = \frac{n_B^{ik}}{n_B^{i\cdot}} \quad , \quad \mathbf{x}_i = \frac{n_A^{i\cdot} + n_B^{i\cdot}}{n_A + n_B} \quad , \tag{7}$$

with $n_A^{i\cdot}$ and $n_B^{i\cdot}$ cardinalities of elements with X_i in samples A and B, respectively, while n_A^{ij} is the cardinality of elements in A with (X_i, Y_j) and n_B^{ik} is the cardinality of elements in B with (X_i, Z_k).

Whenever $n_A^{i\cdot}$ (the same for $n_B^{i\cdot}$) is equal to zero (i.e. no observation in A has X_i) the value $\mathbf{y}_{j|i}$ ($\mathbf{z}_{k|i}$) is undefined and this specific parameter has not any estimation.

If the probabilities $P(S_s), s = 1, 2$, can be elicited, we get a precise conditional probability assessment $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$ with

$$V = \{X_i,\, Y_j,\, Z_k\}\,,\ \mathcal{E} = \{X_i,\, Y_j|X_i,\, Z_k|X_i\}\,,\ \mathbf{p} = \{\mathbf{x}_i, \mathbf{y}_{j|i}, \mathbf{z}_{k|i}\},\ i \in I, j \in J, k \in K, \tag{8}$$

while $\mathfrak{C}$ is a set of logical constraints, in this field named as "structural zeroes", among elements of V.

Usually, the first step is to check the coherence of $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$, that should resort to check the satisfiability of a sequence of linear systems (see, e.g., [5]) but that in the particular context of the statistical matching can be reduced to the solvability of a unique linear system (see [11]). Generally, whenever $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$ is coherent there is more than one solution and the set of all of them forms a so called "credal set".

In the trivial case of logical independence, coherence is automatically ensured (see [11]). In the more worthwhile case of structural zeroes among random variables $\mathcal{Y}$ and $\mathcal{Z}$ (for real applications where these are present refer, e.g., to [9]), coherence of the entire assessment $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$ in (8) is not directly ensured by the separate coherence of the distinct assessments with numerical parts (2), (3), (6). The problem is hence to find a coherent assessment that solves inconsistencies.

Anyhow, whenever present, inconsistencies focus on conditional events with the same conditioning X_i (proofs and examples again in [11]).

This result will permit to split the problem of the merging of the two estimates into separate subproblems: one for the unconditional values $\mathbf{x}_i, i \in I$, and one for each conditioning X_i about the conditional quantities $\{\mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}$, $j \in J$ and $k \in K$. In each of these subproblems the merging and correction procedure can be applied, even being the statistical matching a conditional problem, by fixing in each subproblem the conditioning event, that could be the sure event $\top$ or X_i, and dealing with actually unconditional problems. To see how this could be possible, let us formalize in the next Sections the merging and correction procedure, starting with the formal definition of the unconditional probability assessments.

3 Correction of probability assessments

A probability assessment on a finite domain is a quadruple $\pi = (V, U, p, \mathfrak{C})$, where $V = \{X_1, \ldots, X_k\}$ is a finite set of propositional variables, representing any potential event of interest, U is a subset of V that contains the effective events taken into consideration, $p : U \to [0, 1]$ is a function which assigns a probability value to each variable in U, and $\mathfrak{C}$ is a finite set of logical constraints which lie among all the variables in V.

With such framework, the user provides a probability value for the elements of set U, but logical constraints can also be written in terms of all the existing events V. This feature allows to extend an initial assessment to a larger domain without redefining the whole model.

The constraints in $\mathfrak{C}$ are written with the usual logical notation, where $\neg$, $\wedge$ and $\vee$ denote the negation, disjunction and conjunction connectives, respectively; $\Rightarrow$ the material implication; $=$ the logical equivalence; $\top$ and $\bot$ the universal tautology and contradiction (sure and impossible events), respectively. These constraints can be used to represent any kind of compound event, for instance that an event is the conjunction of other two events, or denote the implications or incompatibilities among the elements of V. Without loss of generality, we suppose that $\mathfrak{C}$ is expressed in conjunctive normal form (CNF) that will help in the implementation part of the correction procedure. Hence $\mathfrak{C} = \{c_1, \ldots, c_m\}$ where each element c_i of $\mathfrak{C}$ is a disjunctive clause, i.e. $c_i = \left(\bigvee_{h \in H_i} X_h \right) \vee \left(\bigvee_{l \in L_i} \neg X_l \right)$ for some $H_i, L_i \subseteq \{1, \ldots, n\}$. Since we will require that all the logical constraint present in $\mathfrak{C}$ must be satisfied, $\mathfrak{C}$ can be seen as the conjunction of $c_1, \ldots, c_m$.

Since a probabilistic assessment π is partial, it may or not be coherent, i.e. consistent with a probability distribution.

The problem of checking the coherence of a probability assessment, called **CPA**, has been already studied (see [3, 4] among the many), albeit in a slightly different form, showing that it is a NP-complete problem, even when the constraints in $\mathfrak{C}$ are binary (i.e., each of them involves only two variables).

There exist several approach to solve **CPA**. Among those, the Mixed Integer Programming (MIP) based approach has proved to be very effective as reported in

Table 1: Variables of $\mathcal{P}1$

name	size	type
a_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$	$n(n+1)$	binary
b_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$	$n(n+1)$	real
q_j, for $j = 1, \ldots, n+1$	$n+1$	real
r_i for $i = 1, \ldots, n$	n	real
s_i for $i = 1, \ldots, n$	n	real

[6, 7], where their implementation was able to handle coherence testing instances up to 1000 variables and 1000 disjunctive clauses in average time ranging from some seconds to some minutes.

When a probability assessment $\pi = (V, U, p, \mathfrak{C})$ is not coherent, then it is possible to "correct" it in order to obtain a coherent probability assessment π' which is as close as possible to π, according to a distance or a pseudo-distance function between probability assessments.

Definition 1 *Given a distance d, a d-correction of a probability assessment $\pi = (V, U, p, \mathfrak{C})$ is a vector p' such that the probability assessment $\pi' = (V, U, p', \mathfrak{C})$ is coherent and $d(p, p')$ is minimized. We denote $\mathcal{C}_d(\pi)$ the sets of all the d-correction of π.*

It is important to notice that for certain choices of d, $\mathcal{C}_d(\pi)$ has just one element, for instance when d is the Euclidean distance. On the other hand, for some other choices of d, $\mathcal{C}_d(\pi)$ has more than one element for some probability assessments π. In this case, the operation of correcting a probability assessment leads to an imprecise probability model, called "credal set". Clearly if π is coherent, then $\mathcal{C}_d(\pi) = \{p\}$, for any distance d of $\mathbb{R}^n$.

In this paper we focus on the L_1 distance defined as $d_1(p, p') = \sum_{i=1}^{n} |p(X_i) - p'(X_i)|$ and we denote $\mathcal{C}_{d_1}(\pi)$ as $\mathcal{C}(\pi)$.

This distance has two important properties. First of all, the correction can be easily interpreted as a cost of changing the probability values, in terms of the sum of the displacements $|p(X_i) - p'(X_i)|$. Minimization of such displacements obeys to the basic principle of minimal change in a numerical uncertainty setting. Secondly, the resulting minimization problem with L_1 distance can be solved by using linear programming with both integer and real variables and this represents a clear computational advantage compared to other distances which require non linear (quadratic, logarithmic, etc.) optimizations tools.

In [1] the details of the MIP-based program $\mathcal{P}1$ implementation have been give. Here we just recall the basic quantities involved in it.

It is well known that if a probability assessmentis coherent, there exists a sparse probability distribution μ so that p' can be written as a convex combination of at most $n+1$ atoms. Let us call $\alpha^{(1)}, \ldots, \alpha^{(n+1)}$ these atoms.

The variables of $\mathcal{P}1$ are summarized in Table 3, while its linear constraints are

$$\sum_{h \in H_i} a_{h,j} + \sum_{l \in L_i} (1 - a_{l,j}) \geq 1 \qquad i = 1, \ldots, m \quad j = 1, \ldots, n+1 \tag{9}$$

$$\sum_{j=1}^{n+1} b_{ij} = p(X_i) + (r_i - s_i) \qquad i = 1, \ldots, n \tag{10}$$

$$0 \leq b_{ij} \leq a_{ij}, \quad a_{ij} - 1 + q_j \leq b_{ij} \leq q_j \qquad i = 1, \ldots, n \quad j = 1, \ldots, n+1 \tag{11}$$

$$\sum_{i=1}^{n+1} q_j = 1 \tag{12}$$

$$r_i \leq 1, \quad s_i \leq 1 \qquad i = 1, \ldots, n \tag{13}$$

The implicit constraint is that all of the variables must be non-negative, as usual in linear programming.

The variables a_{ij} are binary , i.e. constrained in $\{0, 1\}$. Each value a_{ij} should correspond to the atom component $\alpha^{(j)}(X_i)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$. Indeed, the constraint (9) forces each assignment $(a_{1j}, \ldots, a_{nj})$ to satisfy all the clauses $c_i \in \mathfrak{C}$. The values $q_1, \ldots, q_{n+1}$ represent the coefficient of the convex combination which generates p', which also correspond to the probabilities $\mu(\alpha^{(1)}), \ldots, \mu(\alpha^{(n+1)})$. The constraint (11) allows to express the equation

$$b_{ij} = a_{ij} \cdot q_j \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, n+1,$$

without using the multiplication, otherwise $\mathcal{P}1$ would not be a linear problem. Indeed, if $a_{ij} = 0$, then $b_{ij} = 0$ too. On the other hand, if $a_{ij} = 1$, then $a_{ij} - 1 + q_j \leq b_{ij} \leq q_j$ reduces to $q_j \leq b_{ij} \leq q_j$. In this way, for each $i = 1, \ldots, n$ the sum $\sum_{j=1}^{n+1} b_{ij}$ corresponds to $\sum_{j=1}^{n+1} a_{ij} \cdot q_j$. Since $a_{ij} = 1$ if and only if $\alpha^{(j)}$ satisfies X_i, the sum is also equal to $p'(X_i)$.

The variables r_i, s_i are slack variables, which represent, respectively, the positive and the negative difference between $p(X_i)$ and $p'(X_i)$, as implied by the constraint (10). Hence $(r_i - s_i)$ is the correction on the probability of X_i, for each $i = 1, \ldots, n$.

Finally, the objective function to be minimized is

$$\sum_{i=1}^{n} (r_i + s_i) \tag{14}$$

that, being the sum of these corrections, corresponds to the L_1-distance between p and p', i.e., $\sum_{i=1}^{n} |p(X_i) - p'(X_i)|$. Note that for each $i = 1, \ldots, n$, it is impossible that $r_i > 0$ and $s_i > 0$, otherwise the objective function would not be minimized.

It is easy to see that any solution of the linear program $\mathcal{P}1$ corresponds to a L_1-correction p' of p. And vice versa, any L_1-correction p' of p corresponds to a solution of $\mathcal{P}1$.

The optimal value δ for the objective function corresponds to the minimum possible correction on p and any coherent probability assessment $\pi' = (V, U, p', \mathfrak{C})$ such that $d_1(p, p') = \delta$ is a possible solution i.e., p' is an element of $\mathcal{C}(\pi)$. Note that p' can be simply obtained as $p'_i = p_i + r_i - s_i$ for $i = 1, \ldots, n$.

In many situations $\mathcal{C}(\pi)$ has more than one element and the MIP problem is able to find just one solution, which could not be a good representative of all the elements of $\mathcal{C}(\pi)$, as happens when it is an extreme value. Hence program $\mathcal{P}1$ must be associated with an other MIP program $\mathcal{P}2$ to generate all the elements of $\mathcal{C}(\pi)$. In $\mathcal{P}2$ all the constraints and the variables of $\mathcal{P}1$ are reported and it contains a new real variable z, which is subject to the constraints $r_i + s_i \leq z$, for $i = 1, \ldots, n$ (hence $z \geq \max_{i=1,\ldots,n}(r_i + s_i)$), and the new additional constraint $\sum_{i=1}^{n}(r_i + s_i) = \delta$. In this way, the $\mathcal{P}2$ objective function to be minimized is simply z.

The corrected assessment $\bar{\pi} = (V, U, \bar{p}, \mathfrak{C})$ tries to spread the difference δ as much as possible among all the dimensions, i.e. the variables of U. Hence $\bar{p}$ is, in some sense, the most "entropic" point of $\mathcal{C}(\pi)$.

Using $\bar{p}$, it is possible to find the face F_1 of the polytope $\mathcal{Q}$ where $\mathcal{C}(\pi)$ lies. The face F_1 is itself a convex set with at most $n + 1$ atoms as extremal points, which can be found as a part of the solutions of $\mathcal{P}2$ (i.e., the optimal values of a_{ij}).

By looking at the signs of $\bar{p}(X_i) - p(X_i)$, for $i = 1, \ldots, n$, it is also possible to determine the face F_2 of $\mathcal{B}_\pi(\delta)$ which contains $\mathcal{C}(\pi)$. Indeed, F_2 is a convex set with at most n extremal points of the form $p + sign(\bar{p}(X_j) - p(X_j)) \cdot \delta \cdot e_j$.

The whole set of corrections $\mathcal{C}(\pi)$ will result as $F_1 \cap F_2$.

These steps have been implemented in a procedure named Correct that, given in input any partial assessment π, returns the extremal points of the credal set $\mathcal{C}(\pi)$ (for details refer again to [1]).

In Sec.2 we have seen that an incoherent assessment could come by the merging of two separate assessments π_1 and π_2. Let us show how to produce a new coherent probability assessment π_3 which is a "compromise" between π_1 and π_2, keeping as much as possible the information from both.

Depending if the two assessments are compatible (i.e. they give the same values to common variables) or not (i.e. there is an explicit contradiction given by different probabilities to some common variable) there are two different way of defining the joining of them. We report here just the basic notions, referring again to [1] for all the details.

In case of compatibility, it is possible to join directly the two original assessments, so that the merging will result as $\pi_1 \oplus \pi_2 = \text{Correct}(\pi_1 + \pi_2)$. Note that, since such merging procedure is the result of our Correct procedure, its output could be a credal set, as already outlined in the previous Section.

When the probability assessments to be merged are non compatible it is not possible to join directly them into a unique assessment. Hence, in addition to possible initial incoherences present in the separate assessment, we have to tackle with a sure incoherence in the joint one. Anyhow two different correction procedures are possible: a "weighted combination" of the two assessments, or a "assignment to

duplicates". The first approach requires to create a non contradictory probability assessment derived from π_1 and π_2, by choosing a weighted average probability value for each variable in common.

The merging operation between π_1 and π_2 is then defined as the new assessment obtained as correction of the weighted average $\pi_1 \oplus_\omega \pi_2 = \text{Correct}(\pi_1 +_\omega \pi_2)$.

The second approach is to create a probability assessment which maintains both numerical values and to solve the apparent contradiction by adding a new logical variable X_i', for each variable X_i in common. Obviously the logical constraints $\neg X_i \vee X_i'$ and $X_i \vee \neg X_i'$ must be added to $\mathfrak{C} \cup \mathfrak{D}$ to represent the duplicated events $X_i = X_i'$.

Indeed, apart from separate initial incoherences of the two initial assessments π_1 and π_2, the new assessment so obtained $\pi_1 + \pi_2$ is obviously incoherent since the duplicated events with different associated values and the merging operation of π_1 and π_2 results as $\pi_1 \oplus_I \pi_2 = \text{Correct}(\pi_1 + \pi_2)$. Note that, whenever the two assessments π_1 and π_2 are compatible, this merging operator $\pi_1 \oplus_I \pi_2$ coincides with the previous $\pi_1 \oplus \pi_2$ since no duplication of variables is needed in such a case.

The main difference between the two merging of incompatible assessments just described is that $\oplus_I$ is an unsupervised approach since it tries to automatically solve the contradictions, while the operator $\oplus_\omega$ is a supervised approach since it needs an explicit and "exogenous" conciliation among explicit numerical contradictions through the choice of the weight ω. These differences can lead to very different final results. Anyway, the idea behind these two methods is the same, i.e., the merging of two information sources can be performed in two steps. First, put together all the information $\mathcal{I}$, and then find the smallest number of corrections on $\mathcal{I}$ such that the new information $\mathcal{I}'$ is consistent. The choice of which merging operator to adopt should be based on the availability or not of the weight ω representing the relevance, or better of the reliability, of the sources of information. If a reliability grade ω is available, or reasonably assessed, the $\oplus_\omega$ should be preferred, if not the $\oplus_I$ operator avoids the use of unrealistic assumptions.

4 Application of the merging and correction procedures to the statistical matching problem

We can now describe how the merging an correction procedures defined in the previous Section can be applied to the statistical matching problem described in Sec. 2. The preliminary operation is to merge the estimates coming from the two different sampling schemes S_1 and S_2. In particular, since incoherences could be focused only on events conditioned to the same event, we can split the domain $\mathcal{E}$ into sub-domains

$$\mathcal{E}_\Omega = \{X_i\}_{i \in I}\,; \tag{15}$$

$$\mathcal{E}_i = \{Y_j | X_i, Z_k | X_i\}_{j \in J, k \in K} \text{ for } i \in I \tag{16}$$

Since, as described in Section 2, variables $\mathcal{Y}$ and $\mathcal{Z}$ are not jointly observed, on the domains E_i the two sources of information do not overlap and hence the problem will be to, eventually, correct the estimates $\{\mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}$ obtained through (2) and (3). A proper merging operation is needed for the estimates $\{\mathbf{x}_i^{S_1}\}_{i\in I}$ and $\{\mathbf{x}_i^{S_2}\}_{i\in I}$, both on elements of $\mathcal{E}_\Omega$.

As described in Sec. 3, two different approaches can be used: the "supervised" procedure if we can assess the "weight" ω of the relevance or reliability of sources; or the "unsupervised" one that relies on the duplication of all events X_i and consequent addition of structural constraints that express such duplication.

Schematically, the first approach needs hence to compute at first a componentwise"weighted average"

$$\mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2} = \omega\{\mathbf{x}_i^{S_1}\}_{i\in I} + (1-\omega)\{\mathbf{x}_i^{S_2}\}_{i\in I} \tag{17}$$

for a chosen weight $\omega \in [0,1]$, and consequently apply the correct procedure to $(V, \mathcal{E}_\Omega, \mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}, \mathfrak{C})$ obtaining for the numerical part

$$\mathbf{lub} = \mathbf{x}^{S_1} \oplus_\omega \mathbf{x}^{S_2} = \text{Correct}(\mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}) \tag{18}$$

If there is some missing value for $\{\mathbf{x}_i^{S_1}\}_{i\in I}$ or for $\{\mathbf{x}_i^{S_2}\}_{i\in I}$ it must be put equal to 0 in (17). Remember that the correct procedure could lead to either a single solution or to a convex set of solutions, hence **lub** in (18) could be either an actually precise coherent assessment $\{\mathbf{x}_i\}_{i\in I}$ or a proper lower-upper assessment $\{\mathbf{lub}_i\}_{i\in I}$.

Note moreover that, if estimates are taken through frequencies in both samples, $\mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}$ in (17) turns out to be directly coherent for any choice of $\omega \in [0,1]$ so that $\mathbf{lub} = \{\mathbf{x}_i\}_{i\in I} = \mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}$. In particular, choosing $\omega = \frac{n_A}{n_A+n_B}$ we obtain exactly the $\mathbf{x}_i$ estimates already described in (7). So the common sampling scheme can be re-interpreted in our method as separate sampling schemes with weights proportional to the different sample dimensions.

The second approach is to let the correct procedure work without any exogenous weight of the sources and contemplating simultaneously the two different estimates $\{\mathbf{x}_i^{S_1}\}_{i\in I}$ and $\{\mathbf{x}_i^{S_2}\}_{i\in I}$. The obvious inconsistencies are solved by duplicating the events in $\mathcal{E}_\Omega$ as $\mathcal{E}'_\Omega = \{A_i \equiv X_i, B_i \equiv X_i\}_{i\in I}$ and by adding structural zeros induced by the duplicates $A_i = B_i$, for $i \in I$. Hence the correction procedure can be applied to the concatenated assessment $\mathbf{x}^{S_1} \biguplus \mathbf{x}^{S_2}$ that assigns $\mathbf{x}_i^{S_1}$ to A_i and $\mathbf{x}_i^{S_2}$ to B_i, for any $i \in I$, by obtaining a, generally imprecise, assessment $\mathbf{lub} = \mathbf{x}^{S_1} \oplus_I \mathbf{x}^{S_2} = \text{Correct}(\mathbf{x}^{S_1} \biguplus \mathbf{x}^{S_2})$.

As already mentioned, to the other conditioned "strata" $(\mathcal{E}_i, \{\mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}_{j\in J, k\in K})$ the correction procedure can be straightly applied obtaining, generally imprecise, estimates $\{\mathbf{lub}_{j|i}, \mathbf{lub}_{k|i}\}_{j\in J,k\in K}$, for $i \in I$.

At the end, by collecting all the corrections we get a, generally imprecise, coherent assessment $(V, \mathcal{E}, \{\mathbf{lub}_i, \mathbf{lub}_{j|i}, \mathbf{lub}_{k|i}\}_{i\in I, j\in J, k\in K}, \mathfrak{C})$ as the merging of the separate estimates based on the two sample schemes S_1 and S_2.

5 Correction of a statistical matching with misclassification

In [8] it is described a variation of the usual statistical matching problem by introducing a missclassification mechanism that could be summarized by saying that the common variable $\mathcal{X}$ is biasedly observed in source A (e.g. if its values are assessed by not experts in the field) giving rise to a new variable $\mathcal{X}^*$ with the same modalities $mx_i, i \in I$, while $\mathcal{X}$ remains properly observed in the second source B.

In addition, a missclassification mechanism, specified by conditional probabilities $P_{\mathcal{X}|X^*_{i^*}}(X_i)$, can be fully or partially assessed. Hence the whole assessment that results from the joining of all the available information will be of the form $\pi = (V^*, \mathcal{E}^*, \mathbf{p}^*, \mathfrak{C}^*)$ with

$$\begin{aligned} V^* = \{X_i,\, X^*_{i^*}\, Y_j,\, Z_k\} \quad &, \quad \mathcal{E}^* = \{X_i,\, X^*_{i^*}, Y_j | X^*_{i^*},\, Z_k | X_i,\, X_i | X^*_{i^*}\}\,, \\ \mathbf{p}^* = \{\mathbf{x}_i, \mathbf{x}^*_{i^*}, \mathbf{y}_{j|i^*}, \mathbf{z}_{k|i}, \mathbf{x}_{i|i^*}\} \quad &, \quad (i, i^*) \in \mathcal{I} \subseteq I \times I, j \in J, k \in K, \end{aligned} \tag{19}$$

while $\mathfrak{C}^*$ incorporates the structural zeroes among elements of V^*.

This brakes the division in the subdomains (15,16) and the possibility to correct incoherence of the whole assessments with a finite set of corrections on the subdomains. Anyhow, always in [8], it has been proven that the coherence of the whole assessments is basically due to the coherence of the subassessment involving only $\mathcal{X}^*$ and $\mathcal{X}$, hence with numerical part $\mathbf{p}^*_{|\mathcal{I}} = \{\mathbf{x}_i, \mathbf{x}_{i^*}, \mathbf{x}_{i|i^*}\}_{(i,i^*)\in\mathcal{I}}$, and that, in the case of absence of structural zeroes between $\mathcal{X}^*$ and $\mathcal{X}$ (i.e. $\mathcal{I} = I \times I$), the conditional probabilities $\mathbf{x}_{i|i^*}$, $i, i^* \in I$, are constrained by coherence to lay inside the so called Fréchet-Hoeffding bounds:

$$\frac{\max(0, \mathbf{x}_i + \mathbf{x}^*_{i^*} - 1)}{\mathbf{x}^*_{i^*}} \leq \mathbf{x}_{i|i^*} \leq \frac{\min(\mathbf{x}_i, \mathbf{x}^*_{i^*})}{\mathbf{x}^*_{i^*}}. \tag{20}$$

Such bounds imply that the set of coherent values for $\{\mathbf{x}_i, \mathbf{x}_{i^*}, \mathbf{x}_{i|i^*}\}_{(i,i^*)\in\mathcal{I}}$ is not convex in general, hence the credal set of a correction of an incoherent assessments could result not connected and hardly computable. Hence we cannot expect a procedure that produces the whole credal set of correction $\mathcal{C}(\pi)$. Anyhow, we can find just one element of such credal set by a particular setting of linear constraints in a new MIP-based optimization.

More precisely, we change a little bit the notation with respect the MIP program $\mathcal{P}1$ described in Sec.3. In fact now the atoms are characterized by the simple possibility of having the conjunction $X_i \wedge X^*_{i^*}$, so that the set of constraints associated to the subassessment can be simply represented by set of couples of indexes $\mathfrak{C}^*_{|\mathcal{I}} = \{(i, i^*) \in I \times I : X_i \wedge X^*_{i^*} = \bot\}$ (in the sequel we will denote with c^* the cardinality of $\mathfrak{C}^*_{|\mathcal{I}}$). Consequently the binary variables can be denoted with a_{ii^*}, while the real variables with b_{ii^*} and q_{ii^*}. About the slack variables, we need them for the potential modification of both the marginal and conditional probabilities, hence we denote them with $r_i, s_i, r_{i^*}, s_{i^*}, r_{i|i^*}, s_{i|i^*}$, respectively. With such a choice

the constraints of a new MIP program $\mathcal{P}3$ become:

$$\sum_{(i,i^*)\in\mathfrak{C}^*_{|\mathcal{I}}} (1 - a_{ii^*}) \geq c^* \tag{21}$$

$$0 \leq b_{ii^*} \leq a_{ii^*} \qquad a_{ii^*} - 1 + q_{ii^*} \leq b_{ii^*} \leq q_{ii^*} \tag{22}$$

$$\sum_{i^*} b_{ii^*} = \mathbf{x}_i + (r_i - s_i) \tag{23}$$

$$\sum_{i} b_{ii^*} = \mathbf{x}^*_{i^*} + (r_{i^*} - s_{i^*}) \tag{24}$$

$$b_{ii^*} = \mathbf{x}_{i|i^*}\mathbf{x}_{i^*} + \mathbf{x}_{i|i^*}(r_{i^*} - s_{i^*}) + \mathbf{x}_{i^*}(r_{i|i^*} - s_{i|i^*}) \tag{25}$$

$$\sum_{i,i^*\in I} b_{ii^*} = 1 \tag{26}$$

$$r_i \leq 1, \quad s_i \leq 1, \quad r_{i^*} \leq 1 \quad , \quad s_{i^*} \leq 1, \quad r_{i|i^*} \leq 1, \quad s_{i|i^*} \leq 1, \tag{27}$$

where the constraint (21) induces the binary variables a_{ii^*} to be 0 for the couples of indexes in $\mathfrak{C}^*_{|\mathcal{I}}$; constraints like (22) are set for all $i, i^* \in I$ and induce equalities $b_{ii^*} = a_{ii^*}q_{ii^*}$ that otherwise will not be linear; constraints like (23) and (24) are set for all the assessed marginal probabilities $\mathbf{x}_i$ and $\mathbf{x}^*_{i^*}$ and permit their correction through the slack variables; constraints like (25) are set for all assessed conditional probabilities $\mathbf{x}_{i|i^*}$ and constraint the joint distribution with corrected conditional and marginal values. Note that these last type of constraints are equivalent to set

$$b_{ii^*} = (\mathbf{x}_{i|i^*} + r_{i|i^*} - s_{i|i^*})(\mathbf{x}_{i^*} + r_{i^*} - s_{i^*}) \tag{28}$$

but without developing the cross products among the slack variables since they will constitute corrections of the joint distribution that, not being assessed, does not need any correction. This permits us to remain in a linear program.

The objective function to minimize is again the sum of the slack variables

$$\sum_{i,i^*} r_i + s_i + r_{i^*} + s_{i^*} + r_{i|i^*} + s_{i|i^*} \tag{29}$$

that obviously represents the L_1 distance between the assess probability values and the coherent ones.

Note that whenever the corrected assessment would present some marginal probability to be zero, all the new probabilities conditioned on such $X^*_{i^*}$ will result automatically coherent since, for the structure of the assessment, the various zero layers (for such a notion refer to [5]) will involve only one such conditioning event per time, so that the $P_{\mathcal{X}|X^*_{i^*}}(X_i)$ do not have any particular constraint to satisfy.

At the moment we have developed only the theoretical part of this section, leaving its practical application to future developments.

References

[1] M. Baioletti, A. Capotorti. Efficient L_1-Based Probability Assessments Correction: Algorithms and Applications to Belief Merging and Revision. in: *ISIPTA'15 Proc. of the 9th International Symposium on Imprecise Probability: Theories and Applications*. Pescara (IT), 37–46, ARACNE, 2015.

[2] M. Baioletti, A. Capotorti. An efficient probabilistic merging procedure applied to statistical matching. *Lecture Notes in Computer Science*, 10351 LNCS, 65–74, 2017.

[3] Baioletti, M., Capotorti, A., Tulipani, S., Vantaggi B. Simplification Rules for the Coherent Probability Assessment Problem. *Ann. of Math. and Artif. Intell.*, 35:11–28, 2002.

[4] Baioletti, M., Capotorti, A., Tulipani, S. An empirical complexity study for a 2CPA solver. In: *Modern Information Processing: From Theory to Applications*. B. Bouchon-Meunier, G. Coletti and R.R. Yager, Eds. 1–12, 2005.

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

[6] Cozman, F.G., Fargonidi Ianni, L. Probabilistic Satisfiability and Coherence Checking through Integer Programming. Lecture Notes in Computer Science, 7958, 145–156, 2013.

[7] Cozman, F.G., Fargonidi Ianni, L. Probabilistic satisfiability and coherence checking through integer programming. International Journal of Approximate Reasoning, 58, 57–70, 2015.

[8] M. Di Zio, B. Vantaggi. Partial identification in statistical matching with misclassification, *International Journal of Approximate Reasoning*, 82: 227–241, 2017.

[9] D'Orazio, M., Di Zio, M., Scanu, M. Statistical Matching: Theory and Practice, Wiley, New York, 2006.

[10] D'Orazio, M., Di Zio, M., Scanu, M. Statistical Matching for Categorical Data: displaying uncertainty and using logical constraints. Journal of Official Statistics 22, 137–157, 2006.

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

Detecting Correlation Between Extreme Probability Events

Giulianella Coletti
Dept. Mathematics and Computer Science
University of Perugia, Italy
giulianella.coletti@unipg.it

Linda C. van der Gaag
Dept. Information and Computing Sciences
University of Utrecht, NL
l.c.vandergaag@uu.nl

Davide Petturiti
Dept. Economics
University of Perugia, Italy
davide.petturiti@unipg.it

Abstract

Since the classical definitions of correlation give rise to counterintuitive findings for extreme probability events, we build upon the concept of coherent conditional probability to introduce enhanced notions of correlation. Our new notions allow handling extreme events in a principled way by accommodating the different levels of strength of the zero probabilities involved. Where the detection of correlations by means of these levels is computationally challenging, we provide a full characterisation of the correlations between extreme probability events without reference to the complex structure of probability.

1 Introduction

The importance of handling *extreme probability events* in a principled way has been stressed in a range of papers (see for example, [3, 5, 7]); by an extreme probability event we mean a highly unexpected event, that is, an event of zero probability, or a nearly sure event, of probability 1. Zero probabilities necessarily arise in uncountable algebras and, hence, in real-world applications involving infinite settings, where the lack of expressive power of the real numbers often forces possible events to be assigned zero probability. Yet, also in finite settings do extreme probabilities arise. When extracting (conditional) probabilities from real-world data, for example, unexpected events and events occurring with negligible frequency will receive zero probabilities. To forestall the inclusion of zero probabilities in probabilistic models, various more or less "ad hoc" solutions are in use, such as the well-known Laplace correction and the use of pseudocounts in a Bayesian setting. Forcing all

distinguished events to have positive probability however, drastically restricts the class of admissible distributions and, hence, the possibilities of extending partial assessments to complete probabilities.

In applications of probability theory, stochastic independence and the concepts of positive and negative correlation play an important role. While in the context of extreme probabilities stochastic independence has been well studied (see for example [3, 5, 7]), and has led to an enhanced definition of independence, the concept of correlation has received little to no attention. In this paper, we demonstrate that in the presence of extreme probability events, the classical definition of correlation can give counterintuitive results, such as an event E being uncorrelated with an event H logically implying it. Based on these observations, we introduce enhanced notions of correlation which accommodate the different levels of *strength* of the zero probabilities involved. We develop the notions of positive and negative correlation in a coherent setting, referring to full conditional probabilities represented by their complete agreeing classes which in turn define the zero layers of the events of interest. Although the framework of coherent setting constitutes the principle on which our enhanced notion of correlation is founded, referring to zero layers does not provide for practicable application in real-world settings, as a consequence of the computational challenges involved. We therefore provide also a full characterisation of the correlations involving extreme probability events without reference to the complex structure of probability.

The paper is organised as follows. Section 2 presents some preliminaries on coherent conditional probability and thereby introduces our notational conventions. In Section 3, we present our concepts of positive and negative correlation in a coherent setting and introduce some of their properties. Section 4 then provides the characterisation of all correlations involving extreme probability events. Section 5 concludes the paper with our plans for further research.

2 Preliminaries

We consider an *event* to be any fact described by a Boolean sentence, indicating by Ω the *sure event* and using $\varnothing$ for the *impossible event*; for any event E, we will use E^* to indicate either E itself or its contrary E^c. A *conditional event* $E \mid H$ is an ordered pair of events E, H with $H \neq \varnothing$; in the pair, the two events E and H have the same type, both being Boolean sentences, yet have different roles in the sense that H has the role of hypothesis. We recall that an *additive class* of events is a set of events closed under disjunction $\vee$; a *Boolean algebra* of events is an additive class which is further closed under taking the contrary $(\cdot)^c$, and hence under conjunction $\wedge$. For any Boolean algebra $\mathcal{A}$, we use $\mathcal{A}^0$ to denote $\mathcal{A}\backslash\{\varnothing\}$. For an arbitrary family of events $\mathcal{E}$, we use *algebra*$(\mathcal{E})$ to denote the minimal Boolean algebra of events containing $\mathcal{E}$ and *additive*$(\mathcal{E})$ to denote the minimal additive class of events containing $\mathcal{E}$; by *atoms*$(\mathcal{E})$ we indicate the finest partition of Ω contained in *algebra*$(\mathcal{E})$. We will restrict our further discussion to finite Boolean algebras.

In this paper, we build on the following axiomatic definition of *conditional probability* which dates back to de Finetti [8], and has been explicitly formulated, with minor differences, by Dubins [9] and Krauss [10].

Definition 1. *Let $\mathcal{A}$ be a Boolean algebra of events and let $\mathcal{H}$ be an additive class with $\mathcal{H} \subseteq \mathcal{A}^0$. A **conditional probability** on $\mathcal{A}\times\mathcal{H}$ is a function $P\colon \mathcal{A}\times\mathcal{H} \to [0,1]$ that satisfies the following conditions:*

(i) $P(E\,|\,H) = P(E \wedge H\,|\,H)$, *for every* $E \in \mathcal{A}$ *and* $H \in \mathcal{H}$;

(ii) $P(\cdot\,|\,H)$ *is a finitely additive probability on* $\mathcal{A}$, *for every* $H \in \mathcal{H}$;

(iii) $P(E\wedge F\,|\,H) = P(E\,|\,H)\cdot P(F\,|\,E\wedge H)$, *for every* $H, E\wedge H \in \mathcal{H}$ *and* $E, F \in \mathcal{A}$.

Whenever $\Omega \in \mathcal{H}$, we write $P(E) = P(E\,|\,\Omega)$, for every $E \in \mathcal{A}$. Following Dubins, we say that a conditional probability $P(\cdot\,|\,\cdot)$ is *full on* $\mathcal{A}$ if it is defined on $\mathcal{A} \times \mathcal{A}^0$, that is, if $\mathcal{H} = \mathcal{A}^0$. Dubins has shown that every conditional probability on $\mathcal{A} \times \mathcal{H}$ with $\mathcal{H} \subset \mathcal{A}^0$ can be extended to a full conditional probability on $\mathcal{A} \times \mathcal{A}^0$ [9].

For any Boolean algebra of events $\mathcal{A}$, every full conditional probability $P(\cdot\,|\,\cdot)$ on $\mathcal{A}$ has a one-to-one correspondence with a linearly ordered class $\{P_0, \ldots, P_k\}$ of (unconditional) probabilities on $\mathcal{A}$, called its *complete agreeing class*, whose supports form a partition of Ω. For a given full conditional probability $P(\cdot\,|\,\cdot)$, its class $\{P_0, \ldots, P_k\}$ is obtained by setting

- $P_0(\cdot) = P(\cdot\,|\,H_0^0)$, with $H_0^0 = \Omega$;
- for each successive α, $P_\alpha(\cdot) = P(\cdot\,|\,H_0^\alpha)$, with $H_0^\alpha = \bigvee_{H \subseteq H_0^{\alpha-1}, P_{\alpha-1}(H)=0} H \neq \varnothing$;

with the iterative construction halting when $H_0^{k+1} = \varnothing$. We note that for every event $H \in \mathcal{A}^0$, there is an index $\alpha \in \{0, \ldots, k\}$ with $P_\alpha(H) > 0$. Moreover, for every conditional event $E \mid H \in \mathcal{A} \times \mathcal{A}^0$ and α_H being the minimum index in $\{0, \ldots, k\}$ with $P_{\alpha_H}(H) > 0$, we have that

$$P(E\,|\,H) = \frac{P_{\alpha_H}(E \wedge H)}{P_{\alpha_H}(H)}.$$

Having so far addressed full conditional probabilities on an algebra $\mathcal{A}$, we now consider arbitrary, possibly partially specified, conditional probabilities.

Definition 2. *Let $\mathcal{G} = \{E_j\,|\,H_j\}_{j\in J}$, with J a finite index set, be an arbitrary family of conditional events. A **coherent conditional probability** on $\mathcal{G}$ is a function $P\colon \mathcal{G} \to [0,1]$ for which there exists a conditional probability $P'\colon \mathcal{A} \times \mathcal{H} \to [0,1]$, with $\mathcal{A} = \mathit{algebra}(\{E_j, H_j\}_{j\in J})$ and $\mathcal{H} = \mathit{additive}(\{H_j\}_{j\in J})$, such that $P'_{|\mathcal{G}} = P$.*

We note that, since every conditional probability P' on $\mathcal{A}\times\mathcal{H}$ can be extended to a full conditional probability on $\mathcal{A}$, Definition 2 can also be formulated by requiring the existence of a full conditional probability on $\mathcal{A}$ extending the original function P. In the sequel, we will use the phrase *assessment* to denote a function P for

which coherence has yet to be established. The following theorem now specifies several characterisations of coherence for such an assessment, relevant to our current context; for proofs of the equivalences stated in the theorem, we refer to [1, 2, 4].

Theorem 1. *Let $\mathcal{G} = \{E_j \mid H_j\}_{j\in J}$, with J a finite index set, be an arbitrary family of conditional events. Then, for any function $P\colon \mathcal{G} \to [0,1]$, the following statements are equivalent:*

(*i*) *P is a coherent conditional probability;*

(*ii*) *There exists a complete agreeing class $\{P_0, \ldots, P_k\}$, $k \geq 0$, of probabilities P_α on algebra$(\{E_j, H_j\}_{j\in J})$ such that, for every $j \in J$, if α_j is the minimum index in $\{0, \ldots, k\}$ with $P_{\alpha_j}(H_j) > 0$, then*

$$P(E_j \,|\, H_j) = \frac{P_{\alpha_j}(E_j \wedge H_j)}{P_{\alpha_j}(H_j)};$$

(*iii*) *With the atom sets $\mathcal{C}_0 = atoms(\{E_j, H_j\}_{j\in J})$ and, for $\alpha = 1, \ldots, k$, $\mathcal{C}_\alpha = \{C_r \in \mathcal{C}_{\alpha-1} \mid P_{\alpha-1}(C_r) = 0\}$, all systems of equations $\mathcal{S}_\alpha$ in the sequence of systems $\{\mathcal{S}_0, \ldots, \mathcal{S}_k\}$, $k \geq 0$, with non-negative unknowns $x_r^\alpha = P_\alpha(C_r)$ for all $C_r \in \mathcal{C}_\alpha$, are compatible:*

$$\mathcal{S}_\alpha\colon \begin{cases} \displaystyle\sum_{C_r\in\mathcal{C}_\alpha, C_r\subseteq E_j\wedge H_j} x_r^\alpha = P(E_j \,|\, H_j) \cdot \sum_{C_r\in\mathcal{C}_\alpha, C_r\subseteq H_j} x_r^\alpha, \text{ for all } j \in J \text{ with } P_{\alpha-1}(H_j) = 0 \\ \displaystyle\sum_{C_r\in\mathcal{C}_\alpha} x_r^\alpha = 1. \end{cases}$$

Of the sequence of systems $\mathcal{S}_0, \ldots, \mathcal{S}_k$ introduced in Theorem 1(*iii*), every sequence of solutions $\{\mathbf{x}^0, \ldots, \mathbf{x}^k\}$ defines a complete agreeing class $\{P_0, \ldots, P_k\}$ on the algebra *algebra*$(\{E_j, H_j\}_{j\in J})$ by setting $P_0(C_r) = x_r^0$ for all $C_r \in \mathcal{C}_0$, and for each successive $\alpha = 1, \ldots, k$, setting

$$P_\alpha(C_r) = 0 \text{ for every } C_r \in \mathcal{C}_0 \setminus \mathcal{C}_\alpha, \quad \text{and} \quad P_\alpha(C_r) = x_r^\alpha \text{ for every } C_r \in \mathcal{C}_\alpha,$$

and then extending each probability P_α by additivity. In turn, the complete agreeing class $\{P_0, \ldots, P_k\}$ described in Theorem 1(*ii*) has a one-to-one correspondence with a full conditional probability $P'(\cdot\,|\,\cdot)$ on *algebra*$(\{E_j, H_j\}_{j\in J})$ extending P.

To conclude our preliminaries, we recall the concept of *zero layer* [6], which naturally arises from the structure of conditional probability described in Theorem 1.

Definition 3. *Let $\mathcal{A}$ be a Boolean algebra of events and let $P(\cdot\,|\,\cdot)$ be a full conditional probability on $\mathcal{A}$ represented by the complete agreeing class $\{P_0, \ldots, P_k\}$ of probabilities on $\mathcal{A}$. For every event $H \in \mathcal{A}^0$, the* **zero layer** *of H with respect to $\{P_0, \ldots, P_k\}$ is the non-negative number*

$$o(H) = \min\{\alpha \in \{0, \ldots, k\} \,:\, P_\alpha(H) > 0\},$$

with the zero layer of the impossible event equal to $o(\varnothing) = +\infty$*. For every conditional event* $E \,|\, H \in \mathcal{A} \times \mathcal{A}^0$*, the* **zero layer** *of* $E \,|\, H$ *with respect to* $\{P_0, \ldots, P_k\}$ *is the non-negative number*

$$o(E \,|\, H) = o(E \wedge H) - o(H).$$

We note that, for any event E with $P(E) = P(E \,|\, \Omega) > 0$, we have that $o(E) = 0$. We further note that $P(E \,|\, H) > 0$ iff $o(E \wedge H) = o(H)$ and hence $o(E \,|\, H) = 0$.

3 Positive and negative correlation

Before defining our enhanced concept of corrrelation, we review the classical definition of correlation between two events, stated in terms of coherence.

Definition 4. *Let* P *be a coherent conditional probability defined on an arbitrary family of events* $\mathcal{G}$ *with* $E, E \,|\, H \in \mathcal{G}$*. Then,*

- E *is* **positively correlated** *with* H *iff* $P(E \,|\, H) > P(E)$*;*
- E *is* **negatively correlated** *with* H *iff* $P(E \,|\, H) < P(E)$*;*
- E *and* H *are* **not correlated** *iff* $P(E \,|\, H) = P(E)$*.*

Various properties of correlation having been formulated for the classical setting, we review in the following proposition some properties through which we will demonstrate the inadequacy of the classical definitions for describing correlation in the presence of extreme probability events.

Proposition 1. *Let* $\mathcal{G}$ *be an arbitrary family of conditional events including* E^*, H^**,* $E^* \,|\, H^*$*. Let* P *be a coherent conditional probability on* $\mathcal{G}$ *such that* $P(E), P(H) \in]0, 1[$*. Then, the following properties hold:*

- (i) *if* E *is positively (or, alternatively: negatively) correlated with* H*, then* E^c *is positively (negatively) correlated with* H^c*;*
- (ii)
  - *if either* $E \wedge H = \varnothing$ *or* $E^c \wedge H^c = \varnothing$*, then* E *is negatively correlated with* H*;*
  - *if either* $E^c \wedge H = \varnothing$ *or* $E \wedge H^c = \varnothing$*, then* E *is positively correlated with* H*;*
- (iii) E *is positively (negatively) correlated with* H *iff* $P(E \,|\, H) > (<) P(E \,|\, H^c)$*.*

Proof. The properties (i) and (iii) follow directly from Definition 1. The first part of property (ii) follows from the observation that $E \wedge H = \varnothing$ implies $P(E \,|\, H) = 0 < P(E)$. As $E^c \wedge H^c = \varnothing$ implies $P(E^c \,|\, H^c) = 0 < P(E^c)$, we have by property (i) that $P(E \,|\, H) < P(E)$. In both cases, therefore, E is negatively correlated with H. The second part of property (ii) follows analogously. □

We note that property (i) of Proposition 1 strictly depends on the premise that the probabilities of E and H are different from 0 and 1. For property (iii), moreover,

the implication $P(E \mid H) > P(E) \Rightarrow P(E \mid H) > P(E \mid H^c)$ holds only when $P(E), P(H) \in]0,1[$, while the reversed implication is universally valid.

To illustrate the inadequacy of the classical definition above for describing correlation in the presence of extreme probability events, we consider an event H with $P(H) = 1$. By Definition 4, this event is not correlated with any other event, as for any event $E \neq H$ we would find that $P(E \mid H) = P(E)$. We would find the exact same result, in fact, also for an event E which logically contradicts H, as we would then have $P(E \mid H) = P(\varnothing \mid H) = P(E) = 0$. Yet, E could clearly not be considered uncorrelated with H. Similarly counterintuitive conclusions are found for an event E which is logically implied by H and for an event E with $P(E) = 0$.

Not all researchers accept Definition 4 as the basic definition of correlation, however, and may argue that the above observations are due to using an inappropriate definition. They may use property (*iii*) of Proposition 1 for the basic definition of correlation instead, that is, use Definition 5 below.

Definition 5. *Let P be a coherent conditional probability defined on an arbitrary family of events $\mathcal{G}$ with $E\,|\,H, E\,|\,H^c \in \mathcal{G}$. Then,*

- *E is* **positively correlated** *with H iff $P(E\,|\,H) > P(E\,|\,H^c)$;*
- *E is* **negatively correlated** *with H iff $P(E\,|\,H) < P(E\,|\,H^c)$;*
- *E and H are* **not correlated** *iff $P(E\,|\,H) = P(E\,|\,H^c)$.*

We note that Definitions 4 and 5 are not equivalent: while Definition 4 implies Definition 5, the reverse does not hold. In fact, by Definition 5, a conditioning event H with $P(H) = 1$ is not necessarily uncorrelated with an event E. Since $P(H^c) = 0$, there is an index $\alpha_H > 0$ such that $P_{\alpha_H}(H) > 0$ and

$$P(E\,|\,H^c) = \frac{P_{\alpha_H}(E \wedge H^c)}{P_{\alpha_H}(H^c)},$$

which, without any further information, can assume any value in $[0,1]$ and hence also values larger, or smaller, than $P(E)$. Yet, also Definition 5 does not capture the full impact of the hypothesis H on the degree of belief in E when $P(E\,|\,H) = P(E\,|\,H^c) = 0$ or $P(E\,|\,H) = P(E\,|\,H^c) = 1$.

From the above considerations, we conclude that, with both definitions, we need to distinguish between different zeroes, depending on their strengths, before concluding that two extreme probability events are uncorrelated. We provide an example to illustrate our conclusion.

Example 1. *Let Ω be the unit square $[0,1]^2$. Let the event E be the Boolean sentence $E = P \vee Q \vee R$ where P, Q, R are the points $P = \left(\frac{3}{4}, \frac{3}{4}\right)$, $Q = \left(\frac{1}{2}, \frac{1}{2}\right)$, $R = \left(\frac{3}{4}, \frac{1}{4}\right)$ in Ω; let the event $H = \{(x,y) \mid x = y,\ x, y \in [0,1]^2\}$ be the diagonal of the unit square. In this setting, we consider the following assessment for a family of four conditional events:*

$$P(E\,|\,H) = P(E\,|\,H^c) = P(H\,|\,E^c) = 0, \quad P(H\,|\,E) = \frac{2}{3}$$

For proving coherence, we consider the set atoms$(\{E,H\}) = \{C_1, C_2, C_3, C_4\}$ *with*

$$C_1 = E \wedge H = P \vee Q, \quad C_2 = E \wedge H^c = R, \quad C_3 = E^c \wedge H, \quad C_4 = E^c \wedge H^c,$$

and build the sequence of systems $\mathcal{S}_\alpha$ *with non-negative unknowns* $x_i^\alpha = P_\alpha(C_i)$*, as described in Theorem 3. The first system equals*

$$\mathcal{S}_0 : \begin{cases} x_1^0 = 0 \cdot (x_1^0 + x_3^0) \\ x_2^0 = 0 \cdot (x_2^0 + x_4^0) \\ x_3^0 = 0 \cdot (x_3^0 + x_4^0) \\ x_1^0 = \frac{2}{3} \cdot (x_1^0 + x_2^0) \\ x_1^0 + x_2^0 + x_3^0 + x_4^0 = 1 \end{cases}$$

which has $x_1^0 = x_2^0 = x_3^0 = 0$, $x_4^0 = 1$*, for its solution. Then, focusing on the zero-probability atoms and writing* x_i^1 *for* x_i^0*, the second system is found to be*

$$\mathcal{S}_1 : \begin{cases} x_1^1 = 0 \cdot (x_1^1 + x_3^1) \\ x_1^1 = \frac{2}{3} \cdot (x_1^1 + x_2^1) \\ x_1^1 + x_2^1 + x_3^1 = 1 \end{cases}$$

which has $x_1^1 = x_2^1 = 0$, $x_3^1 = 1$*, for its unique solution. The third system equals*

$$\mathcal{S}_2 : \begin{cases} x_1^2 = \frac{2}{3} \cdot (x_1^2 + x_2^2) \\ x_1^2 + x_2^2 = 1 \end{cases}$$

and has $x_1^2 = \frac{2}{3}$, $x_2^2 = \frac{1}{3}$ *for its sole solution. Since every constructed system of the sequence has a unique solution, the assessment* P *has a unique complete agreeing class* $\{P_0, P_1, P_2\}$*. This class implies that*

$$o(E\,|\,H) = o(E \wedge H) - o(H) = 2 - 1 \;<\; 2 - 0 = o(E \wedge H^c) - o(H^c) = o(E\,|\,H^c).$$

The zero layer of taking the event H *for the hypothesis thus is smaller than that of taking* H^c *for the hypothesis. As the conditional event* $E \mid H^c$ *still has zero probability in the structure when* $E \mid H$ *does not, this finding may be naturally construed as a positive correlation of* E *and* H*.*

We further consider the incompatible events R *and* H*. Analogously to the above example, we find for the conditional events* $R|H$ *and* $R|H^c$ *that*

$$o(R\,|\,H) = o(R \wedge H) - o(H) = +\infty - 1 > 2 - 0 = o(R \wedge H^c) - o(H^c) = o(R\,|\,H^c),$$

which demonstrates that the logical impossibility of R *under the hypothesis* H *results in a zero layer which is infinitely larger than that of* R *under the hypothesis* H^c*. The zero probability resulting from a logical impossibility will thus always be deeper in the complex structure of probability than the zero probability of any possible event.*

Based on the considerations in the above example, we now introduce our definition of correlation of extreme probability events in a coherent setting.

Definition 6. *Let P be a coherent conditional probability defined on an arbitrary family of conditional events $\mathcal{G}$ containing $E^* \,|\, H^*, H^* \,|\, E^*$. We say that:*

- *E is* **positively correlated in a coherent setting** *with H, denoted as $E \perp_{cs}^{+} H$, if one of the following conditions holds:*
  - $P(E \,|\, H) > P(E \,|\, H^c)$;
  - $P(E \,|\, H) = P(E \,|\, H^c) = 0$, *and every complete agreeing class $\{P_\alpha\}$ on algebra$(\{E, H\})$ that agrees with P on $\mathcal{D}$, has*
  $$o(E \,|\, H) < o(E \,|\, H^c);$$
  - $P(E \,|\, H) = P(E \,|\, H^c) = 1$, *and every complete agreeing class $\{P_\alpha\}$ on algebra$(\{E, H\})$ that agrees with P on $\mathcal{D}$, has*
  $$o(E^c \,|\, H) > o(E^c \,|\, H^c);$$
- *E is* **negatively correlated in a coherent setting** *with H, denoted as $E \perp_{cs}^{-} H$, if one of the following conditions holds:*
  - $P(E \,|\, H) < P(E \,|\, H^c)$;
  - $P(E \,|\, H) = P(E \,|\, H^c) = 0$, *and every complete agreeing class $\{P_\alpha\}$ on algebra$(\{E, H\})$ that agrees with P on $\mathcal{D}$, has*
  $$o(E \,|\, H) > o(E \,|\, H^c),$$
  - $P(E \,|\, H) = P(E \,|\, H^c) = 1$, *and every complete agreeing class $\{P_\alpha\}$ on algebra$(\{E, H\})$ that agrees with P on $\mathcal{D}$, has*
  $$o(E^c|H) < o(E^c|H^c);$$
- *E is* **not correlated in a coherent setting** *with H, denoted as $E \not\perp_{cs} H$, if it is not positively nor negatively correlated in a coherent setting with H.*

The above definition of positive and negative correlation in a coherent setting avoids the counterintuitive findings from the classic definitions of correlation which were illustrated above. In fact, in the presence of extreme probability events, the definition allows the identification of a correlation between events which are logically related, as shown in the following theorem.

Theorem 2. *Let P be a coherent conditional probability defined on an arbitrary family of conditional events $\mathcal{G}$ containing $E^* \mid H^*, H^* \mid E^*$. Then, the following properties hold:*

(*i*) *if either* $E \wedge H = \varnothing$ *or* $E^c \wedge H^c = \varnothing$, *then* $E \perp^-_{cs} H$;

(*ii*) *if either* $E^c \wedge H = \varnothing$ *or* $E \wedge H^c = \varnothing$ *then* $E \perp^+_{cs} H$.

Proof. We prove property (*i*); the proof of property (*ii*) is analogous. If $E \wedge H = \varnothing$, we just have to consider the case where $P(E\,|\,H) = 0 = P(E\,|\,H^c)$; in this case we have that $o(E\,|\,H) = +\infty > o(E\,|\,H^c)$ and the negative correlation follows. Similarly, if $E^c \wedge H^c = \varnothing$, we address just the case where $P(E\,|\,H) = 1 = P(E\,|\,H^c)$; since then $o(E^c\,|\,H) < +\infty = o(E^c\,|\,H^c)$, the negative correlation equally follows. □

We note that the correlations $\perp^+_{cs}$ and $\perp^-_{cs}$ introduced above generally are not symmetric, as demonstrated by the following example.

Example 2. *We address two events* E, H *which are logically independent, and consider the following coherent probability assessment for these events:*

$$P(E\,|\,H) = \frac{3}{4},\ P(E\,|\,H^c) = \frac{1}{4},\ P(E) = \frac{1}{4},\ P(H) = 0.$$

By definition, we have that E *is positively correlated with* H *and, hence,* $E \perp^+_{cs} H$.

We now address the way in which H *is correlated with* E. *Building upon the set of atoms* $atoms(\{E, H\}) = \{C_1, C_2, C_3, C_4\}$ *with*

$$C_1 = E \wedge H,\ C_2 = E \wedge H^c,\ C_3 = E^c \wedge H,\ C_4 = E^c \wedge H^c,$$

we consider the sequence of systems $\mathcal{S}_\alpha$ *with non-negative unknowns* $x^\alpha_r = P_\alpha(C_r)$ *as before. The first system of equations equals*

$$\mathcal{S}_0 : \begin{cases} x^0_1 = \frac{3}{4} \cdot (x^0_1 + x^0_3) \\ x^0_2 = \frac{1}{4} \cdot (x^0_2 + x^0_4) \\ x^0_1 + x^0_2 = \frac{1}{4} \cdot (x^0_1 + x^0_2 + x^0_3 + x^0_4) \\ x^0_1 + x^0_3 = 0 \cdot (x^0_1 + x^0_2 + x^0_3 + x^0_4) \\ x^0_1 + x^0_2 + x^0_3 + x^0_4 = 1 \end{cases}$$

having $x^0_1 = x^0_3 = 0$, $x^0_2 = \frac{1}{4}$, $x^0_4 = \frac{3}{4}$ *for its solution. The second system then is*

$$\mathcal{S}_1 : \begin{cases} x^1_1 = \frac{3}{4} \cdot (x^1_1 + x^1_3) \\ x^1_1 + x^1_3 = 1 \end{cases}$$

whose unique solution is $x^1_1 = \frac{3}{4}$, $x^1_3 = \frac{1}{4}$. *These solutions determine the unique agreeing class* $\{P_0, P_1\}$ *on the algebra* $algebra(\{E, H\})$, *which, in turn, has a one-to-one correspondence with a full conditional probability* P' *on* $algebra(\{E, H\})$ *extending* P. *This probability* P' *has*

$$P'(H\,|\,E) = \frac{P_0(H \wedge E)}{P_0(E)} = 0 = \frac{P_0(H \wedge E^c)}{P_0(E^c)} = P'(H\,|\,E^c),$$

and, hence,

$$o(H\,|\,E) = o(H \wedge E) - o(E) = 1 = o(H \wedge E^c) - o(E^c) = o(H\,|\,E^c),$$

from which we find that $H \not\perp^{+}_{cs} E$. *We conclude that, while* E *is positively correlated with* H, *the reverse does not hold.*

For symmetric concepts of positive and negative correlation in a coherent setting, the definitions of $\perp^{+}_{cs}$ and $\perp^{-}_{cs}$ need be further enhanced, by setting

$$\begin{aligned} E \perp^{+}_{S-cs} H \quad &\text{iff} \quad E \perp^{+}_{cs} H \text{ and } H \perp^{+}_{cs} E, \\ E \perp^{-}_{S-cs} H \quad &\text{iff} \quad E \perp^{-}_{cs} H \text{ and } H \perp^{-}_{cs} E. \end{aligned}$$

Because of space limitations, we do not further elaborate on this enhancement.

4 Detecting correlations in a coherent setting

Detecting correlations in the presence of extreme probability events by means of the definitions introduced in the previous section, involves the construction of a sequence of systems of equations to determine the zero layers of the conditional probabilities involved. The next theorem now characterizes the possible correlations between two logically independent events E and H, in terms of just the probabilities $P(H), P(E^* \mid H^*)$ and $P(H^* \mid E^*)$. The theorem thereby provides for detecting all correlations between the two events without the need to explicitly identify the zero layers for the conditional events involved.

Theorem 3. *Let* E, H *be logically independent events and let* P *be a coherent conditional probability on a family of conditional events* $\mathcal{G}$ *containing the subset* $\mathcal{D} = \{E^*\,|\,H^*, H^*\,|\,E^*\}$, *with* $P(E\,|\,H) = P(E\,|\,H^c)$. *Then, the following properties hold:*

(i) $E \perp^{+}_{cs} H$ *if and only if one of the following conditions holds:*

- (a) $P(E \mid H) = 0$ *and all extensions of* P *to* $H, H \mid E$ *meet either of the following conditions:*
  1. $P(H) = 0$ *and* $P(H\,|\,E) > 0$;
  2. $0 < P(H) < 1$ *and* $P(H\,|\,E) = 1$;
- (b) $P(E\,|\,H) = 1$ *and all extensions of* P *to* H *and* $H|E$ *meet either of the following conditions:*
  1. $P(H) = 0$ *and* $P(H\,|\,E^c) > 0$;
  2. $0 < P(H) < 1$ *and* $P(H\,|\,E^c) = 1$;

(ii) $E \perp^{-}_{cs} H$ *if and only if one of the following conditions holds:*

- (c) $P(E \mid H) = 0$ *and all extensions of* P *to* $H, H \mid E$ *meet either of the following conditions:*

1. $P(H) > 0$ *and* $P(H \,|\, E) = 0$*;*
2. $P(H) = 1$ *and* $0 < P(H \,|\, E) < 1$*;*

(*d*) $P(E \mid H) = 1$ *and all extensions of* P *to* $H, H \mid E$ *meet either of the following conditions:*

1. $P(H) = 0$ *and* $P(H \,|\, E^c) > 0$*;*
2. $P(H) = 1$ *and* $0 < P(H \,|\, E^c) < 1$*.*

Proof. For proving the theorem, we take $atoms(\{E, H\}) = \{C_1, C_2, C_3, C_4\}$ with

$$C_1 = E \wedge H,\ C_2 = E \wedge H^c,\ C_3 = E^c \wedge H,\ C_4 = E^c \wedge H^c.$$

We further consider a complete class $\{P_\alpha\}$ on $algebra(\{E, H\})$ agreeing with the restriction of P to $\mathcal{D}$, obtained by solving a sequence of systems $\mathcal{S}_\alpha$ as in Theorem 1.

We first prove that condition (*a*)*1.* implies property (*i*); proofs of the conditions (*a*)*2.* and (*b*) implying (*i*) are analogous. We assume that $P(E \,|\, H) = P(E \,|\, H^c) = 0$ and, moreover, that $P(H) = 0$ and $P(H \,|\, E) > 0$; we take $P(H \,|\, E) = p \in]0, 1]$. Under these conditions, every complete agreeing class $\{P_\alpha\}$ that agrees with P on $\mathcal{D}$, has $P_0(C_4) = 1$, $P_1(C_3) = 1$, $P_2(C_1) = p$ and $P_2(C_2) = 1 - p$, which implies that $o(E \wedge H) = 2$ and $o(E \wedge H^c) \geq 2$ while $o(H) = 1$ and $o(H^c) = 0$. We conclude that $o(E \,|\, H) < o(E \,|\, H^c)$ and, hence, that $E \perp^+_{cs} H$.

We now prove that condition (*a*) suffices for concluding $E \perp^+_{cs} H$; the proof involving condition (*b*) is analogous. We assume $P(E \,|\, H) = P(E \,|\, H^c) = 0$, $o(E \,|\, H) < o(E \mid H^c)$, and take $P(H \mid E) = q \in]0, 1]$. We now distinguish between three cases:

- We suppose that $P(H) = \delta \in]0, 1[$. The only complete agreeing class satisfying $o(E \mid H) < o(E \mid H^c)$ is the class $\{P_0, P_1, P_2\}$ with $P_0(C_3) = \delta$, $P_0(C_4) = 1 - \delta$, $P_1(C_1) = 1$ and $P_2(C_2) = 1$, which implies that $P(H \mid E) = 1$.

- We suppose that $P(H) = 0$. Every complete agreeing class having $P_0(C_4) = 1$, $P_1(C_1) = 0$, we consider the following possibilities for the remaining atoms:

  - if $P_1(C_2) = \delta \in]0, 1[$ and $P_1(C_3) = 1 - \delta$, we must have that $P(H|E)$ equals zero and $o(E|H) = o(E|H^c)$, which contradicts our assumption;
  - if $P_1(C_3) = 1$ and, hence, $P_2(C_1) = P(H \,|\, E) = q$ and $P_2(C_2) = 1 - q$, it follows that $o(E \,|\, H) = o(E \,|\, H^c)$, which contradicts our assumption;
  - if $P_1(C_2) = 1$, we must have that $P(H \,|\, E) = 0$ and, as $P_2(C_3) = 1$ and $P_3(C_1) = 1$, also $o(E \,|\, H) = o(E \,|\, H^c)$, contradicting our assumption.

- We cannot have $P(H) = 1$, as this would contradict $o(E \,|\, H) < o(E \,|\, H^c)$.

□

5 Concluding observations

Based on the observation that the classical definitions of correlation can give counterintuitive results in the presence of extreme probability events, we provided an enhanced definition of correlation in a coherent setting. To allow ready applicability of our definition to real-world applications, we gave a full characterisation of correlations involving extreme probability events without referring to the underlying complex structure of the probability involved. We noted that our definition of correlation in a coherent setting is not symmetric; as a next step in our research, we will address this asymmetry by studying the conditions under which it occurs. In the future, we will investigate how our enhanced definition of correlation can be embedded in the framework of qualitative probabilistic influence, to render this framework suitable to real-world applications involving extreme probability events.

References

[1] G. Coletti. Coherent numerical and ordinal probabilistic assessments. *IEEE Transactions on Systems, Man and Cybernetics*, 24(0):1747–1754, 1994.

[2] G. Coletti and R. Scozzafava. Characterization of coherent probabilities as a tool for their assessment and extension. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 44:101–132, 1996.

[3] G. Coletti and R. Scozzafava. Null events and stochastical independence. *Kybernetika*, 34:69–78, 1998.

[4] G. Coletti and R. Scozzafava. Conditioning and inference in intelligent systems. *Soft Computing*, 3:118–130, 1999.

[5] G. Coletti and R. Scozzafava. *Zero probabilities in stochastic independence*, volume 516 of *The Springer International Series in Engineering and Computer Science*, pages 185–196. Springer, Boston, 2000.

[6] G. Coletti and R. Scozzafava. Probabilistic logic in a coherent setting. volume 15 of *Trends in Logic*. Kluwer Academic Publishers, Dordrecht, 2002.

[7] G. Coletti and R. Scozzafava. Stochastic independence in a coherent setting. *Annals of Mathematics and Artificial Intelligence*, 35:151–176, 2002.

[8] B. de Finetti. *Probability, Induction and Statistics: The art of guessing.* John Wiley & Sons, London, New York, Sydney, Toronto, 1972.

[9] L. Dubins. Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations. *Annals of Probability*, 3(1):89–99, 1975.

[10] P. Krauss. Representation of conditional probability measures on boolean algebras. *Acta Mathematica Academiae Scientiarum Hungarica*, 19(3-4):229–241, 1968.

Decisions on generalized Anscombe-Aumann acts under possibly "unexpected" scenarios

Giulianella Coletti
Dept. Mathematics and Computer Science
University of Perugia, Italy
giulianella.coletti@unipg.it

Davide Petturiti
Dept. Economics
University of Perugia, Italy
davide.petturiti@unipg.it

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

Abstract

We consider decisions on generalized Anscombe-Aumann acts, mapping states of the world to belief functions over a set of consequences. Preference relations on these acts are given by a decision maker under different scenarios (conditioning events). Then, we provide a system of axioms which are necessary and sufficient for the representability of these "conditional preferences" through a conditional functional $\mathbf{CEU}_{P,u}$, parametrized by a unique full conditional probability P on the algebra of events and a cardinal utility function u on consequences. The model is able to manage also "unexpected" (i.e., "null") conditioning events. We finally provide an elicitation procedure that reduces to a Quadratically Constrained Linear Problem (QCLP).

1 Introduction

In many decision problems under uncertainty in economics, we need to choose between uncertain consequences in a set X that are contingent on the states of the world in S. So, we distinguish between an "objective" uncertainty related to X (i.e., exogenously quantified and given to the decision maker, in the spirit of von Neumann-Mergenstern) and a "subjective" uncertainty related to S (i.e., encoded in the decision maker's preferences, in the spirit of Savage). This configures a two-stage process where first the state of the world is chosen by Nature, and then the consequence is chosen through "objective" uncertainty, in the spirit of [1].

Very often, due to partial knowledge, uncertainty cannot be encoded in a single probability measure, but we rather have a class of probability measures.

We refer to situations where ambiguity is related to the "objective" probabilistic assessment as that due to a partially known randomizing device (like an urn or a roulette wheel) that results in a class of probability measures whose lower envelope is a *belief function* [10, 27], like in the well-known Ellsberg's urn paradox [12]. Following [28], in these cases we will speak of "objective" ambiguity. Hence, the above objects of decisions can be modelled as generalized Anscombe-Aumann acts [1] mapping S to the set $\mathbf{B}(X)$ of belief functions over X, forming the set $\mathcal{F} = \mathbf{B}(X)^S$.

A crucial aspect of making decisions under uncertainty is the possibility of reasoning under hypotheses. Unexpected situations such as earthquakes, terror attacks or financial crises are normally identified with "null" events and are often ignored in decision problems. Nevertheless, "unexpected" scenarios can deeply impact on the analysis of a decision problem [19] and should not be discarded.

Here we consider a conditional decision model involving the above generalization of Anscombe-Aumann acts, assuming that the decision maker is able to provide a family of preference relations $\{\precsim_H\}_{H\in\wp(S)^0}$ on $\mathcal{F}$ indexed by the set $\wp(S)^0 = \wp(S)\setminus\{\emptyset\}$ of non-impossible events. Every preference relation $\precsim_H$ can be interpreted as comparing acts under the hypothesis H.

In the model we propose, "objective" ambiguity is expressed by referring to the class of belief functions over X (as in the models [4, 17]). On the other hand, "subjective" uncertainty is assumed to be probabilistic, so, we model it with a *full conditional probability* in the sense of [8, 11, 23], that allows for conditioning to "null" events, but possible.

Here, we search for a representation in terms of a conditional functional $\mathbf{CEU}_{P,u}$ parametrized by a full conditional probability $P(\cdot|\cdot)$ on $\wp(S)\times\wp(S)^0$ and a utility function $u : X \to \mathbb{R}$. The above conditional functional consists in a mixture with respect to a full conditional probability of Choquet expected utilities [4] contingent on the states of the world. In particular, due to the properties of the Choquet integral [25], every state-contingent Choquet expected utility is actually a lower expected utility with respect to the probabilities in $\mathbf{core}(f(s))$. The present model generalizes the conditional version of the Anscombe-Aumann model given in [21] by introducing "objective" ambiguity.

We provide a set of axioms for the family $\{\precsim_H\}_{H\in\wp(S)^0}$ that is proved to be necessary and sufficient for the existence of a unique full conditional probability $P(\cdot|\cdot)$ and a cardinal utility function u such that the corresponding $\mathbf{CEU}_{P,u}$ functional *represents* the preferences, i.e., for every $f, g \in \mathcal{F}$ and every $H \in \wp(S)^0$,

$$f \precsim_H g \iff \mathbf{CEU}_{P,u}(f|H) \leq \mathbf{CEU}_{P,u}(g|H).$$

It turns out that a rational agent in this model behaves as a $\mathbf{CEU}_{P,u}$ maximizer, so, as a maximizer of a conditional expected value of state-contingent lower expected utilities. Hence, the present model encodes a form of "objective" ambiguity aversion. The model can be easily extended in a way to cope with different attitudes

towards "objective" ambiguity: this will be the subject of future research.

A similar decision setting, limited to the unconditional case, has been considered by [28], where the author takes acts mapping states of the world to non-empty compact convex polyhedral sets of probability measures over consequences. In the same paper the author considers a representation functional different from ours, but still relying on a mixture with respect to a "subjective" probability measure.

Important efforts have been addressed in the decision theory literature to model "subjective" ambiguity, that is to ambiguity in "subjective" uncertainty evaluations (see, e.g., the survey papers [13] and [15]). For instance, in the seminal papers [26] and [16], the classical Anscombe-Aumann setting is considered but there ambiguity is "subjective", since the mixture of state-contingent expected utilities is done through the Choquet integral with respect to a capacity over S in the first model, while a class of "subjective" probabilities is considered in the second model. Still working in the classical Anscombe-Aumann setting, we find the models [2, 3, 20]. Other lines of research take care of "subjective" ambiguity in a Savage's setting, through acts that map states of the world to non-empty sets of consequences [14, 22]. All the quoted decision models essentially focus on unconditional decisions.

The conditional functional $\mathbf{CEU}_{P,u}$ is completely specified once the full conditional probability $P(\cdot|\cdot)$ and the utility function u have been elicited by the decision maker. In general, an agent is only able to provide few comparisons for few conditioning events. In this case, the first issue is to check the consistency of the given comparisons with the model of reference. When consistency holds, it is easily seen that an elicitation procedure relying on a finite number of arbitrary comparisons cannot guarantee the uniqueness of P and the cardinality of u in general.

We provide an elicitation procedure that reduces to a Quadratically Constrained Linear Problem (QCLP). Unfortunately, the quadratic constraints in the problem are generally not positive definite, so, the problem is generally not convex: interior points algorithms are not suitable. The problem can be solved with a branch and bound algorithm coping with global optimization of non-linear problems, such as the Couenne optimizer [7].

2 Model description

Consider the following decision-theoretic setting:

- $X = \{x_1, \ldots, x_m\}$, a finite set of consequences;
- $\wp(X)^0 = \wp(X) \setminus \{\emptyset\}$, the set of multi-consequences, i.e., non-empty sets of consequences;
- $\mathbf{B}(X) = \{Bel : \wp(X) \to [0,1]\}$, the set of all belief functions on $\wp(X)$;
- $S = \{s_1, \ldots, s_n\}$, a finite set of states of the world;
- $\wp(S)$, the set of events;

- $\wp(S)^0 = \wp(S) \setminus \{\emptyset\}$, the set of scenarios, i.e., non-impossible events;
- $\mathcal{F} = \mathbf{B}(X)^S = \{f : S \to \mathbf{B}(X)\}$, the set of all acts;
- $\{\precsim_H\}_{H \in \wp(S)^0}$, a family of preference relation on $\mathcal{F}$, indexed by the set of non-impossible events $H \in \wp(S)^0$.

For every $H \in \wp(S)^0$, we denote with $\prec_H$ and $\sim_H$ the asymmetric and symmetric parts of $\precsim_H$. Moreover, for every $f, g \in \mathcal{F}$, $f \precsim_H g$ means "f is not preferred to g under the hypothesis H", $f \prec_H g$ means "g is preferred to f under the hypothesis H", and $f \sim_H g$ means "f is indifferent to g under the hypothesis H".

Notice that the set $\mathbf{B}(X)$ contains the set

$$\mathbf{B}_0(X) = \{\delta_B \ : \ B \in \wp(X)^0\},$$

of *vacuous belief functions*, where δ_B is the belief function whose Möbius inversion is such that $m_{\delta_B}(B) = 1$ and 0 otherwise. Let us notice that $\mathbf{B}(X)$ is closed with respect to the *convex combination* operation defined, for every $Bel_1, Bel_2 \in \mathbf{B}(X)$ and every $\alpha \in [0, 1]$, pointiwise, for every $A \in \wp(X)$, as

$$(\alpha Bel_1 + (1 - \alpha)Bel_2)(A) = \alpha Bel_1(A) + (1 - \alpha)Bel_2(A),$$

and it holds

$$m_{\alpha Bel_1 + (1-\alpha)Bel_2} = \alpha m_{Bel_1} + (1 - \alpha)m_{Bel_2}.$$

The set of acts $\mathcal{F}$ contains, in particular, the set of *constant acts* $\mathcal{F}_c$ whose elements are defined, for every $Bel \in \mathbf{B}(X)$, as

$$\overline{Bel}(s) = Bel, \ \forall s \in S.$$

The set $\mathcal{F}$ is closed with respect to the following operation of *convex combination*: for every $f, g \in \mathcal{F}$ and every $\alpha \in [0, 1]$, $\alpha f + (1 - \alpha)g$ is defined pointwise, for every $s \in S$, as

$$(\alpha f + (1 - \alpha)g)(s) = \alpha f(s) + (1 - \alpha)g(s).$$

For every $H \in \wp(S)^0$, the relation $\precsim_H$ determines a relation $\trianglelefteq_H$ on $\mathbf{B}(X)$ through constant acts defined, for every $Bel_1, Bel_2 \in \mathbf{B}(X)$, as

$$Bel_1 \trianglelefteq_H Bel_2 \iff \overline{Bel_1} \precsim_H \overline{Bel_2}.$$

In turn, the relation $\trianglelefteq_H$ determines a relation $\leq^\bullet_H$ on $\wp(X)^0$ defined, for every $A, B \in \wp(X)^0$, as

$$A \leq^\bullet_H B \iff \overline{\delta_A} \trianglelefteq_H \overline{\delta_B}.$$

Finally, the relation $\leq^\bullet_H$ induces a relation $\leq^*_H$ on X defined, for every $x, y \in X$, as

$$x \leq^*_H y \iff \{x\} \leq^\bullet_H \{y\}.$$

Let $\leq^*$ be a weak order on X with asymmetric and symmetric parts $<^*$ and $=^*$, respectively, and assume $x_{\sigma(1)} \leq^* \ldots \leq^* x_{\sigma(m)}$, where σ is a permutation

of $\{1,\dots,m\}$. Then, denote $X^* = X_{/=^*} = \{[x_{i_1}],\dots,[x_{i_t}]\}$ for which $<^*$ is a strict order, and we can assume $[x_{i_1}] <^* \cdots <^* [x_{i_t}]$. The $\leq^*$-*aggregated Möbius inversion* associated to $Bel \in \mathbf{B}(X)$ is the function $M^{\leq^*}_{Bel} : X^* \to [0,1]$ defined, for every $[x_{i_j}] \in X^*$, as

$$M^{\leq^*}_{Bel}([x_{i_j}]) = \sum_{x_i \in [x_{i_j}]} \sum_{x_i \in B \subseteq E^\sigma_i} m_{Bel}(B), \tag{1}$$

where $E^\sigma_i = \{x_{\sigma(i)},\dots,x_{\sigma(m)}\}$ for $i = 1,\dots,m$. Note that $M^{\leq^*}_{Bel}([x_{i_j}]) \geq 0$ for every $[x_{i_j}] \in X^*$ and $\sum_{j=1}^t M^{\leq^*}_{Bel}([x_{i_j}]) = 1$, thus $M^{\leq^*}_{Bel}$ determines a probability distribution on X^*. It is easily seen that, if $u : X \to \mathbb{R}$ then defining $x \leq^* y$ if and only if $u(x) \leq u(y)$, for every $Bel \in \mathbf{B}(X)$, it holds

$$\oint u \mathrm{d}Bel = \sum_{[x_{i_j}] \in X^*} u(x_{i_j}) M^{\leq^*}_{Bel}([x_{i_j}]).$$

Let us stress that $M^{\leq^*}_{Bel}$ encodes a pessimistic aggregation of the uncertainty expressed by m_{Bel} [4]. Indeed, it holds

$$\sum_{[x_{i_j}] \in X^*} u(x_{i_j}) M^{\leq^*}_{Bel}([x_{i_j}]) = \sum_{B \in \wp(X)^0} \left(\min_{x \in B} u(x) \right) m_{Bel}(B).$$

We are searching for a representation of $\{\precsim_H\}_{H \in \wp(S)^0}$ in the form of a conditional mixture of Choquet integrals, i.e., for every $f \in \mathcal{F}$ and $H \in \wp(S)^0$,

$$\mathbf{CEU}_{P,u}(f|H) = \sum_{s \in S} P(\{s\}|H) \left(\oint u \mathrm{d}f(s) \right), \tag{2}$$

where $P(\cdot|\cdot)$ is a full conditional probability on $\wp(S) \times \wp(S)^0$ and $u : X \to \mathbb{R}$ is a cardinal utility function.

Consider the following axioms.

(AA1C) Weak order: $\forall H \in \wp(S)^0$, $\precsim_H$ is a weak order on $\mathcal{F}$;

(AA2C) Continuity: $\forall H \in \wp(S)^0$, $\forall f,g,h \in \mathcal{F}$, if $f \prec_H g \prec_H h$, $\exists \alpha,\beta \in (0,1)$ such that

$$\alpha f + (1-\alpha)h \prec_H g \prec_H \beta f + (1-\beta)h;$$

(AA3C) Independence: $\forall H \in \wp(S)^0$, $\forall f,g,h \in \mathcal{F}$ and $\forall \alpha \in (0,1)$

$$f \precsim_H g \iff \alpha f + (1-\alpha)h \precsim_H \alpha g + (1-\alpha)h;$$

(AA4C) Monotonicity: $\forall H \in \wp(S)^0$, $\forall f,g \in \mathcal{F}$, if $f(s) \trianglelefteq_H g(s)$, $\forall s \in S$ then $f \precsim_H g$;

(AA5C) Non-triviality: $\forall H \in \wp(S)^0$, $\exists f, g \in \mathcal{F}$ such that $f \prec_H g$;

(AA6C) Relevance: $\forall H \in \wp(S)^0$, $\forall f, g \in \mathcal{F}$ with $f(s) = g(s)$, $\forall s \in H$ then $f \sim_H g$;

(AA7C) Uncertainty independence: $\forall f, g \in \mathcal{F}$ and $\forall H, K \in \wp(S)^0$, if $f \precsim_H g$, $f \precsim_K g$, and $H \cap K = \emptyset$ then $f \precsim_{H \cup K} g$;

(AA8C) State neutrality: $\forall s, t \in S$, if $f(s) = f(t)$, $g(s) = g(t)$, and $f \precsim_{\{s\}} g$ then $f \precsim_{\{t\}} g$.

(AA9C) Aggregate indifference: $\forall H \in \wp(S)^0$, $\forall A \in \wp(S)$ and $\forall f, g \in \mathcal{F}$ with $f(s) = g(s)$ $\forall s \in A$, if $M_{f(s)}^{\leq_H^*} = M_{g(s)}^{\leq_H^*}$ $\forall s \in A^c$ then $f \sim_H g$;

Axioms **(AA1C)**–**(AA5C)** are the usual Anscombe-Aumann axioms in the formulation of [26], stated for generalized Anscombe-Aumann acts and every preference relation in $\{\precsim_H\}_{H \in \wp(S)^0}$. Axioms **(AA6C)**–**(AA8C)** cope with conditioning. In particular, axiom **(AA6C)** expresses a focusing conditioning rule, i.e., it states that in conditioning to H, only the part of acts inside of H counts. Axiom **(AA7C)** copes with relating different conditioning events, while axiom **(AA8C)** encodes a form of consistency between different states. Finally, axiom **(AA9C)** is responsible for the $\mathbf{CEU}_{P,u}$ representation: it says that if two possibly distinct acts have the same $\leq_H^*$-aggregated Möbius inversion (i.e., the same pessimistic aggregation of "objective" uncertainty) then, they should be judged indifferent given H.

The following theorem, whose proof is omitted due to a lack of space, shows that axioms **(AA1C)**–**(AA9C)** are necessary and sufficient to get a $\mathbf{CEU}_{P,u}$ representation.

Theorem 1. *The following statements are equivalent:*

(i) the family of relations $\{\precsim_H\}_{H \in \wp(S)^0}$ *satisfies* **(AA1C)**–**(AA9C)***;*

(ii) there exist a full conditional probability $P : \wp(S) \times \wp(S)^0 \to [0, 1]$ *and a non-constant utility function* $u : \wp(X)^0 \to \mathbb{R}$ *such that, for every* $f, g \in \mathcal{F}$ *and every* $H \in \wp(S)^0$, $f \precsim_H g \iff \mathbf{CEU}_{P,u}(f|H) \leq \mathbf{CEU}_{P,u}(g|H)$.

Moreover, P *is unique and* u *is unique up to positive linear transformations.*

Let us stress that a $\mathbf{CEU}_{P,v}$ functional allows to take "null" (possible) conditioning events as hypotheses and, even more, it allows to order events in $\wp(S)^0$ according to their "unexpectation". For that, we define, for every $H, K \in \wp(S)^0$,

$$H \sqsubseteq K \iff \mathbf{1}_\emptyset \prec_{H \cup K} \mathbf{1}_H,$$

with the meaning "H is no more unexpected than K", where the act $\mathbf{1}_E$, for $E \in \wp(S)$, is defined as in the proof of Theorem 1. The statement $H \sqsubseteq K$ expresses the uncertainty evaluation $P(H|H \cup K) > 0$, i.e., it considers the probability of the

events H under the hypothesis that either H or K is true. In particular, $H \sqsubset K$ means $P(H|H \cup K) > 0$ and $P(K|H \cup K) = 0$, whereas $H =^{\square} K$ stands for $P(H|H \cup K) > 0$ and $P(K|H \cup K) > 0$. The relation $\sqsubseteq$ reveals to be a weak order on $\wp(S)^0$ and has been originally introduced by [9, 18, 24].

Every full conditional probability $P(\cdot|\cdot)$ on $\wp(S)$ is in bijection with a linearly ordered class of probability measure $\{P_0, \ldots, P_k\}$ on $\wp(S)$, said *complete agreeing class*, whose supports form a partition of S [5, 6].

Events with probability 0 essentially determine the structure of a full conditional probability $P(\cdot|\cdot)$ on $\wp(S)$ and actually the relation $\sqsubseteq$ is intimately related to $\{P_0, \ldots, P_k\}$.

Given $P(\cdot|\cdot)$, the corresponding complete agreeing class $\{P_0, \ldots, P_k\}$ representing it can be built through the events

$$H_0^\alpha = \{s \in H_0^{\alpha-1} \,:\, P(\{s\}|H_0^{\alpha-1}) = 0\} \quad \text{for } \alpha = 1, \ldots, k,$$

with $H_0^0 = S$, by setting $P_\alpha(\cdot) = P(\cdot|H_0^\alpha)$ with $H_0^\alpha \neq \emptyset$. On the other hand, given $\{P_0, \ldots, P_k\}$, for every $E|H \in \wp(S) \times \wp(S)^0$ there is a minimum index $\alpha_H \in \{0, \ldots, k\}$ such that $P_{\alpha_H}(H) > 0$ and it holds

$$P(E|H) = \frac{P_{\alpha_H}(E \cap H)}{P_{\alpha_H}(H)}.$$

The class of events $\{H_0^0, \ldots, H_0^k\}$ determines a decreasing class $\{\mathcal{I}_0, \ldots, \mathcal{I}_k\}$ of ideals of $\wp(S)$, singled out by the relation $\sqsubseteq$, defined as

$$\mathcal{I}_\alpha = \{A \in \wp(S)^0 \,:\, H_0^\alpha \sqsubseteq A\} \cup \{\emptyset\} = \{A \in \wp(S) \,:\, A \subseteq H_0^\alpha\}.$$

The class of events $\{H_0^0, \ldots, H_0^k\}$ also gives rise to a partition $\mathcal{E} = \{E_0, \ldots, E_k\}$ of S obtained by setting

$$E_\alpha = H_0^\alpha \setminus H_0^{\alpha-1} \quad \text{for } \alpha = 0, \ldots, k-1,$$

with $E_k = H_0^k$, where $E_\alpha = \mathbf{supp}(P_\alpha) = \{s \in S \,:\, P_\alpha(\{s\}) > 0\}$ in the complete agreeing class representing $P(\cdot|\cdot)$.

3 Model elicitation

The conditional functional $\mathbf{CEU}_{P,u}$ is completely specified once the full conditional probability $P(\cdot|\cdot)$ and the utility function u have been elicited by the decision maker (DM). In general, the DM is only able to provide few comparisons for few conditioning events. In this case, the first issue is to check the consistency of the given comparisons with the model of reference. When consistency holds, it is easily seen that an elicitation procedure relying on a finite number of arbitrary comparisons cannot guarantee the uniqueness of P and u in general.

Fixed X and S, we propose an elicitation procedure based on three different cognitive tasks.

We ask the DM to determine a subset $\mathcal{L} = \{H_1, \ldots, H_N\} \subseteq \wp(S)$ that correspond to those events considered as "scenarios of interest" and then to order them according to their unexpectation, by providing a weak order $\sqsubseteq$ on $\mathcal{L}$.

We ask the DM to provide a weak order $\leq^*$ on X, i.e., on consequences obtained with certainty.

For every $H \in \mathcal{L}$, we ask the DM to provide a finite number of strict $\{f_l \prec_H g_l\}_{l \in L_H}$ and weak comparisons $\{f_w \precsim_H g_w\}_{w \in W_H}$, with $L_H \neq \emptyset$ while W_H is allowed to be empty. This assures non-triviality.

The issue is to find a complete agreeing class $\{P_0, \ldots, P_k\}$ on $\wp(S)$ (and, so, a full conditional probability $P(\cdot|\cdot)$) *compatible* with the relation $\sqsubseteq$ on $\mathcal{L}$ (that is such that $H_i \sqsubseteq H_j \iff P(H_i|H_i \cup H_j) > 0$) and a utility function $u : X \to \mathbb{R}$ increasing with respect to $\leq^*$, such that the corresponding $\mathbf{CEU}_{P,u}$ conditional functional preserves all the strict and weak preference comparisons.

At this aim, let $\mathcal{L}_{/=^\square} = \{[H_{i_1}], \ldots, [H_{i_M}]\}$ and assume $[H_{i_1}] \sqsubset \ldots \sqsubset [H_{i_M}]$. Now, define $B_0^{M+1} = \emptyset$ and for $\alpha = 0, \ldots, M$, $B_0^\alpha = \bigcup_{\beta=\alpha}^M \bigcup_{H \in [H_{i_\beta}]} H$ and $E_0^\alpha = B_0^\alpha \setminus B_0^{\alpha+1}$.

Every linearly ordered class of probability measures $\{P_0^*, \ldots, P_M^*\}$ on $\wp(S)$ where $\mathbf{supp}(P_\alpha^*) \subseteq E_0^\alpha$, for $\alpha = 0, \ldots, M$, is said *minimal agreeing class* and determines a conditional probability $P^*(\cdot|\cdot)$ on $\wp(S) \times \mathbf{add}(\mathcal{L})$, where $\mathbf{add}(\mathcal{L})$ is the set of events obtained closing $\mathcal{L}$ with respect to unions. The conditional probability $P^*(\cdot|\cdot)$ can be further extended (generally not in a unique way) to a full conditional probability $P(\cdot|\cdot)$ on $\wp(S)$ *compatible* with $\sqsubseteq$ on $\mathcal{L}$. One of the possible extensions is determined by the *complete agreeing class* $\{P_0^*, \ldots, P_M^*, P_{M+1}^*\}$ where P_{M+1}^* is an arbitrary probability measure on $\wp(S)$ such that $\mathbf{supp}(P_{M+1}^*) = S \setminus \bigcup_{\alpha=0}^M \mathbf{supp}(P_\alpha^*)$. The adjunct of P_{M+1}^* is necessary only if $S \setminus \bigcup_{\alpha=0}^M \mathbf{supp}(P_\alpha^*) \neq \emptyset$.

With such an input, the elicitation procedure consists in solving the following optimization problem with unknowns the minimal agreeing class $\{P_0^*, \ldots, P_M^*\}$, the utility function u and the dummy variable δ:

$$\text{maximize } \delta \text{ subject to:}$$

$$\begin{cases} \displaystyle\sum_{s \in E_0^\alpha} P_\alpha^*(\{s\}) \left(\sum_{[x_{i_j}] \in X^*} u(x_{i_j}) \left(M_{f_l(s)}^{\leq^*}([x_{i_j}]) - M_{g_l(s)}^{\leq^*}([x_{i_j}]) \right) \right) + \delta \leq 0, \\ \displaystyle\sum_{s \in E_0^\alpha} P_\alpha^*(\{s\}) \left(\sum_{[x_{i_j}] \in X^*} u(x_{i_j}) \left(M_{f_w(s)}^{\leq^*}([x_{i_j}]) - M_{g_w(s)}^{\leq^*}([x_{i_j}]) \right) \right) \leq 0, \\ \displaystyle\sum_{s \in E_0^\alpha} P_\alpha^*(\{s\}) = 1, \\ P_\alpha^*(\{s\}) \geq 0, \ \forall s \in E_0^\alpha, \\ u(x_{i_1}) = 0, \ u(x_{i_t}) = 1, \ u(x_{i_j}) - u(x_{i_{j+1}}) + \delta \leq 0, \text{ for } j = 1, \ldots, t-1, \\ -1 \leq \delta \leq 1, \end{cases}$$

for $\alpha = 1, \ldots, M$ and all $H \in [H_{i_\alpha}]$, for all $l \in L_H$, for all $w \in W_H$. The above optimization problem is a Quadratically Constrained Linear Problem (QCLP) that is a particular case of a Quadratically Constrained Quadratic Problem (QCQP). Unfortunately, the quadratic constraints in the problem are generally not positive definite, so, the problem is generally not convex: interior points algorithms are not suitable. The problem can be solved with a branch and bound algorithm coping with global optimization of non-linear problems, such as the COUENNE optimizer [7].

Solving the above optimization problem allows to check both the consistency of the given preference statements and, if consistency holds, to find a full conditional probability $P(\cdot|\cdot)$ and a utility function u determining the conditional functional $\mathbf{CEU}_{P,u}$. Indeed, the preference statements are consistent with the model if and only if $\delta > 0$ and in this case the solution of the system determines $P(\cdot|\cdot)$ and u, up to the possible arbitrary choice of the probability measure P^*_{M+1}.

4 A paradigmatic example

Take the set of states of the world $S = \{s_1, s_2, s_3, s_4\}$ spanned by events

- K = "North Korea and USA enter into war next year";
- G = "Italian GDP increases next year";

with $K = \{s_1, s_2\}$ and $G = \{s_1, s_3\}$.

Consider three unitary financial instruments that can result in a loss of €50, in a null gain or in a gain of €100, implying $X = \{-50, 0, 100\}$. From statistics of previous years we only have partial information on the performances of each instrument, that are listed below:

Instrument 1: It is only known that it guarantees a gain of €100 in 30% of cases;

Instrument 2: It is only known that it results in a loss of €50 in 20% of cases;

Instrument 3: No information is available.

Hence, instrument i determines a class of probability measures $\mathbf{P}^i$ on $\wp(X)$ whose lower envelope is easily shown to be a belief function $Bel_i = \min \mathbf{P}^i$, with:

$$\begin{aligned}
\mathbf{P}^1 &= \{P : \wp(X) \to [0,1] | P \text{ is a probability measure}, \gamma \in [0, 0.7], \\
&\qquad P(\{-50\}) = \gamma, P(\{0\}) = 0.7 - \gamma, P(\{100\}) = 0.3\}, \\
\mathbf{P}^2 &= \{P : \wp(X) \to [0,1] | P \text{ is a probability measure}, \gamma \in [0, 0.8] \\
&\qquad P(\{-50\}) = 0.2, P(\{0\}) = \gamma, P(\{100\}) = 0.8 - \gamma\}, \\
\mathbf{P}^3 &= \{P : \wp(X) \to [0,1] | P \text{ is a probability measure}\}.
\end{aligned}$$

Consider the following investment strategies in which the adopted financial instrument is contingent on the state of the world:

	s_1	s_2	s_3	s_4
f	Bel_3	Bel_1	Bel_1	Bel_2
g	Bel_3	Bel_3	Bel_2	Bel_3

The question is: How should a DM decide between f and g conditionally to events K and K^c?

Suppose that our DM is not able to express directly his preference between f and g, conditionally to K and K^c. Nevertheless, our DM is a profit maximizer and believes that a war between North Korea and USA next year is unexpected, while it is more likely a decrease of Italian GDP next year.

The fact that event K is unexpected, i.e., it is judged as "null" by our DM, does not rule out its possible realization. In particular, if event K were true then our DM believes that it would be more likely an increase of Italian GDP, due to a profit of Italian weapons factories.

Hence, our DM is able to provide the following information: $\mathcal{L} = \{K, K^c\}$ with $K^c \sqsubset K$; $-50 <^* 0 <^* 100$; the worst and best multi-consequences $\underline{A} = \{-50\}$ and $\overline{A} = \{100\}$. For every $E \in \wp(S)$, define the act

$$\mathbf{1}_E(s) = \begin{cases} \delta_{\overline{A}}, & \text{if } s \in E, \\ \delta_{\underline{A}}, & \text{if } s \notin E. \end{cases}$$

In turn, the beliefs of our DM can be translated as follows:

$$\mathbf{1}_{\{s_3\}} \prec_{K^c} \mathbf{1}_{\{s_4\}} \quad \text{and} \quad \mathbf{1}_{\{s_2\}} \prec_K \mathbf{1}_{\{s_1\}}.$$

In this case we have that $E_0^0 = K^c$ and $E_0^1 = K$. To avoid cumbersome notation, denote $p_i^\alpha = P_\alpha^*(\{s_i\})$ and $u_1 = u(-50), u_2 = u(0), u_3 = u(100)$. We need to solve the following optimization problem

$$\text{maximize } \delta \text{ subject to:}$$
$$\begin{cases} -p_3^0 u_1 + p_3^0 u_3 + p_4^0 u_1 - p_4^0 u_3 + \delta \le 0, \\ p_1^1 u_1 - p_1^1 u_3 - p_2^1 u_1 + p_2^1 u_3 + \delta \le 0, \\ p_3^0 + p_4^0 = 1, \\ p_3^0, p_4^0 \ge 0, \\ p_1^1 + p_2^1 = 1, \\ p_1^1, p_2^1 \ge 0, \\ u_1 = 0,\ u_2 = 1,\ u_1 - u_2 + \delta \le 0,\ u_2 - u_3 + \delta \le 0, \\ -1 \le \delta \le 1, \end{cases}$$

for which the COUENNE optimizer finds the solution $p_3^0 = 0.18358$, $p_4^0 = 0.81642$, $p_1^1 = 0.81642$, $p_2^1 = 0.18358$, $u_1 = 0$, $u_2 = 0.5$, $u_3 = 1$, and $\delta = 0.5$. Since $\delta > 0$ the preference statements are consistent with the model and a full conditional probability $P(\cdot|\cdot)$ on $\wp(S)$ is that represented by the complete agreeing class $\{P_0^*, P_1^*\}$ whose distributions are

	$\{s_1\}$	$\{s_2\}$	$\{s_3\}$	$\{s_4\}$
P_0^*	0	0	0.18358	0.81642
P_1^*	0.81642	0.18358	0	0

With such $P(\cdot|\cdot)$ and u we have

$$\begin{aligned} \mathbf{CEU}_{P,u}(g|K) = 0 &< 0.055074 = \mathbf{CEU}_{P,u}(f|K), \\ \mathbf{CEU}_{P,u}(g|K^c) = 0.073432 &< 0.381642 = \mathbf{CEU}_{P,u}(f|K^c), \end{aligned}$$

so, $g \prec_K f$ and $g \prec_{K^c} f$, i.e., under both hypothesis the DM should choose f.

References

[1] F. Anscombe and R. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 1963.

[2] P. Casaca, A. Chateauneuf, and J. Faro. Ignorance and competence in choices under uncertainty. *Journal of Mathematical Economics*, 54:143–150, 2014.

[3] A. Chateauneuf and J. Faro. Ambiguity through confidence functions. *Journal of Mathematical Economics*, 45(9):535–558, 2009.

[4] G. Coletti, D. Petturiti, and B. Vantaggi. Rationality principles for preferences on belief functions. *Kybernetika*, 51(3):486–507, 2015.

[5] G. Coletti and R. Scozzafava. Characterization of coherent conditional probabilities as a tool for their assessment and extension. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 44:101–132, 1996.

[6] G. Coletti and R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, volume 15 of *Trends in Logic*. Kluwer Academic Publisher, Dordrecht/Boston/London, 2002.

[7] Couenne. https://projects.coin-or.org/Couenne/.

[8] B. de Finetti. Sull'impostazione assiomatica del calcolo delle probabilità. *Annali Triestini*, 19:29–81, 1949.

[9] B. de Finetti. *Theory of Probability 1-2*. John Wiley & Sons, London, New York, Sydney, Toronto, 1975.

[10] A. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[11] L. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3(1):89–99, 1975.

[12] D. Ellsberg. Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.

[13] J. Etner, M. Jeleva, and J.-M. Tallon. Decision Theory under Ambiguity. *Journal of Economic Surveys*, 26(2):234–270.

[14] P. Ghirardato. Coping with ignorance: unforeseen contingencies and non-additive uncertainty. *Economic Theory*, 17:247–276, 2001.

[15] I. Gilboa and M. Marinacci. *Ambiguity and the Bayesian Paradigm*, chapter 21. Springer International Publishing Switzerland 2016.

[16] I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

[17] J.-Y. Jaffray. Linear utility theory for belief functions. *Operations Research Letters*, 8(2):107–112, 1989.

[18] P. Krauss. Representation of conditional probability measures on Boolean algebras. *Acta Mathematica Hungarica*, 19(3):229–241, 1968.

[19] D. Kreps and R. Wilson. Sequential equilibria. *Econometrica*, 50(4):863–894, 1982.

[20] F. Maccheroni, M. Marinacci, and A. Rustichini. Ambiguity Aversion, Robustness, and the Variational Representation of Preferences. *Econometrica*, 74(6):1447–1498.

[21] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[22] K. Nehring. Preference for flexibility in a Savage framework. *Econometrica*, 67:101–119, 1999.

[23] A. Rényi. On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungarica*, 6(3):285–335, 1955.

[24] A. Rényi. On conditional probability spaces generated by a dimensionally ordered set of measures. *Theory of Probability & Its Applications*, 1(1):55–64, 1956.

[25] D. Schmeidler. Integral representation without additivity. *Proceedings of the American Mathematical Society*, 97(2):255–261, 1986.

[26] D. Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.

[27] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[28] M.-L. Vierø. Exactly what happens after the Anscombe–Aumann race? *Economic Theory*, 41(2):175–212, 2009.

About Two Consonant Conflicts of Belief Functions

Milan Daniel
Jan Becher – Karlovarská Becherovka, a.s.
Pernod Ricard group
milan.daniel@pernod-ricard.com

Václav Kratochvíl
Inst. of Inf. Theory and Automation
Czech Academy Sciences
velorex@utia.cas.cz

Abstract

General belief functions usually bear some internal conflict which comes mainly from disjoint focal elements. Analogously, there is often some conflict between two (or more) belief functions. After the recent observation of hidden conflicts (seminar CJS'17 [17]), appearing at belief functions with disjoint focal elements, importance of interest in conflict of belief functions has increased.

This theoretical contribution introduces a new approach to conflicts (of belief functions). Conflicts are considered independently of any combination rule and of any distance measure.

Consonant conflicts are based on consonant approximations of belief functions in general; two special cases of the consonant approach based on consonant inverse pignistic and consonant inverse plausibility transforms are discussed.

Basic properties of the newly defined conflicts are presented, analyzed and briefly compared with our original approaches to conflict (combinational conflict, plausibility conflict and comparative conflict), with the recent conflict based on non-conflicting parts, as well as with W. Liu's degree of conflict.

1 Introduction

Belief functions (BFs; introduced in [25]) are one of the widely used formalisms for uncertainty representation and processing - that enables representation of incomplete and uncertain knowledge, belief updating, and combination of evidence.

Complications with highly conflicting belief functions combination, see e.g., [9, 28], have motivated a theoretical investigation of conflicts between belief functions [2, 4, 11, 18, 21, 22, 23, 24]. The problematic issue of an essence of conflict between belief functions - originally defined by the non-normalised version of Dempster's rule $\textcircled{\cap}$ (i.e., by its value for the empty set: $m_{\textcircled{\cap}}(\emptyset)$) - was first mentioned by Almond [1], and discussed further by W. Liu [22]. Almond's counter-example has

been overcome by W. Liu's progressive approach. Unfortunately, the substance of the issue has not been solved there as positive conflict still may be detected for non-conflicting BFs.

Further steps ahead were presented in our previous study [11] where new ideas concerning interpretation, definition, and measurement of conflicts of BFs were introduced. Three new approaches to interpretation and computation of conflicts were suggested: combinational conflict, plausibility conflict (see also [13, 14]), and comparative conflict; pignistic conflict analogous to plausibility one was defined later in [14]. Unfortunately, none of those captures the nature of conflict sufficiently enough ands these approaches need further elaboration. Nevertheless, the very important distinction between conflict of two BFs and the internal conflict of an individual BF was pointed out in [11] - altogether with the necessity to distinguish between a conflict and a difference/distance of two BFs; this was also pointed out in [3].

Probabilistic approximations of belief functions were used in several previous approaches, e.g. pignistic transform in W. Liu's two-dimensional degree of conflict [22] and in pignistic conflict [14], normalized plausibility of singletons in plausibility conflict [11, 13, 14], etc.

Unfortunately, application of a probability approximation adds a new additional information, which increases internal conflict of inputs and also resulting in a global conflict. The new reverse approach suggested in this paper adds no new information but removes an information creating the internal conflicts, as inverse probabilistic transformations are used to make consonant approximations. This is an analogy to belief discounting, but without necessity of any parameter due to its specific context. Thus BFs without internal conflicts are used for a computation of a conflict; it is a generalization of the approach from [15] in fact.

2 Preliminaries

2.1 General Primer on Belief Functions

We assume classic definitions of basic notions from theory of *belief functions* [25] on a finite frame of discernment $\Omega_n = \{\omega_1, \omega_2, ..., \omega_n\}$. 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 - then we speak about *normalized bba*. 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* $Pl : \mathcal{P}(\Omega) \longrightarrow [0, 1]$, $Pl(A) = \sum_{\emptyset \neq A \cap X} m(X)$. There is a unique correspondence among m and corresponding Bel and Pl thus we often speak about m as about belief function.

A *focal element* is a subset X of the frame of discernment, such that $m(X) > 0$. Let $\mathcal{F} = \{X \mid m(X) > 0\}$ be the set of all focal elements; and *core* be its union $\mathcal{C} = \bigcup_{X \in \mathcal{F}} X$. If all the focal elements are *singletons* (i.e. one-element subsets of Ω), then we speak about a *Bayesian belief function* (BBF). If all the focal elements are either singletons or whole Ω (i.e. $|X| = 1$ or $|X| = |\Omega|$), then we speak

about a *quasi-Bayesian belief function* (qBBF). If all focal elements have non-empty intersections, we call this a *consistent belief function*. And if all focal elements are nested, we call this a *consonant belief function*. *Vacuous BF* (VBF) has the only focal element Ω: $m_{vac}(\Omega) = 1$. A *symmetric BF* is a BF, which has the same bbms for focal elements with the same cardinality, i.e., $m(X) = m(Y)$ for $|X| = |Y|$.

Let us recall *normalized plausibility of singletons*[1] of Bel: the BBF (probability distribution) $Pl\_P(Bel)$ such, that $(Pl\_P(Bel))(\omega_i) = \frac{Pl(\{\omega_i\})}{\sum_{\omega\in\Omega} Pl(\{\omega\})}$ [5, 10]; and alternative Smets' *pignistic probability*[2] $BetP(\omega_i) = \sum_{\omega_i\in X} \frac{m(X)}{|X|}$ [27].

2.2 A Graphical Presentation of Sets of Belief Functions

We can represent any BF on an n-element frame of discernment Ω_n by an enumeration of its m values (bbms), i.e., by a $(2^n–2)$-tuple $(x_1, x_2, ..., x_{2^n-2})$ as $m(\emptyset) = 0$ and $m(\Omega) = x_{2^n-1} = 1 - \sum_{i=1}^{2^n-2} x_i$. Thus we can present set of all BFs on Ω_n by a $(2^n - 2)$-dimensional simplex in general. Specially we have 2D triangle and 6D simplex for Ω_2 and Ω_3, see Figure 1 [19, 20] and Figure 2 [12].

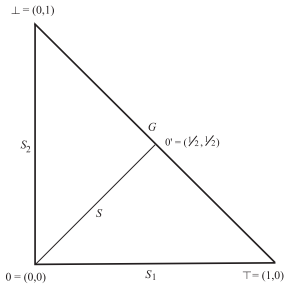


Figure 1: Belief functions on 2-element frame Ω_2; G: Bayesian BFs, S: symmetric BFs, S_1, S_2: simple support BFs ($\sim$ consonant BFs on Ω_2).

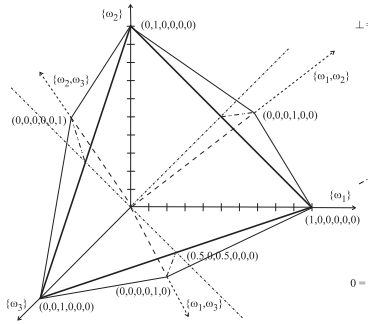


Figure 2: Simplex of Belief functions on 3-element frame Ω_3. 6 dimensions corresponds to 6 possible focal elements.

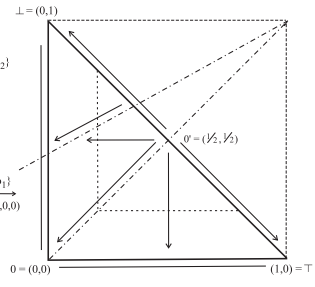


Figure 3: Internal conflict $Pl\text{-}IntC$ on Ω_2. It has max value $\frac{1}{2}$ for $0'$, decreases along arrows, constant along lines without arrows; zero at S_i's.

3 Conflicts of Belief Functions

Conflicts of belief functions are caused mainly by disjoint focal elements either within individual BFs or in different BFs. *Internal conflicts* $IntC(m_i)$ of individual BFs are distinguished from *conflict between BFs* $Conf(m_1, m_2)$ [11]; the entire sum of multiples of mutually conflicting masses is called *total conflict* $TotC(m_1, m_2)$[3].

[1] $Pl\_P(Bel)$ is a normalization of contour function (of plausibility of singletons [25]) in fact.

[2] We have to note an analogy between pignistic probability and Shapley value [26].

[3] Some authors (see e.g. [18]) use 'total conflict' for maximal possible conflict, which arises when all focal elements of one BF are disjoint with focal element of another BFs. Our total

3.1 Internal Conflict of Belief Functions

Internal conflict of a BF is caused either by its disjoint focal elements (if there are any), or if $Pl(\{\omega\}) < 1$ for every $\omega \in \Omega$ (i.e. $\forall\omega \in \Omega$ exists focal element X_ω such that *omega* $\notin X_\omega$). Let us accept the following simple definition of internal conflict[4]: *Internal conflict of BF Bel* is defined by formula $IntC(Bel) = 1 - max_{\omega\in\Omega} Pl(\{\omega\})$, where Pl is the plausibility corresponding to Bel. A BF Bel is *(internally) non-conflicting* when it has zero internal conflict $IntC(Bel) = 0$; it is *(internally) conflicting* otherwise. This definition corresponds to internal plausibility conflict $Pl\text{-}IntC(Bel)$ from [11], see Section 4.1.

Thus a BF is non-conflicting if and only if there is an $\omega \in \Omega$ such that $Pl(\{\omega\}) = 1$ (or in other words if BF is consistent).

3.2 Conflicts between Belief Functions

There are several different assumptions about conflicts between belief functions in our previous approaches [11, 13, 14, 15]. Some of them are mutually conflicting as coming from various alternative approaches, thus we suppose only those common:

A1. Non-negativity and boundary conditions: $0 \leq Conf(Bel_1, Bel_2) \leq 1$.
A2. Symmetry: $Conf(Bel_1, Bel_2) = Conf(Bel_2, Bel_1)$.
A3. $Conf(Bel, Bel) = 0$. A BF is not conflicting with itself.
A4. $Conf(Bel, Bel_{vac}) = 0$. Vacuous BF is non-conflicting with any other BF.

The other assumptions in our previous approaches are stronger and they distinguish among various approaches. Thus we do not consider them among our general assumptions here. We may compare our assumptions with Martin's axioms MA1 – MA5 [23] and Destercke & Burger properties P3 – P6 [18]:

(MA1) : $Conf(Bel', Bel'') \geq 0$,
(MA2) : $Conf(Bel, Bel) = 0$,
(MA3) : $Conf(Bel', Bel'') = Conf(Bel'', Bel')$,
(MA4) : $Conf(Bel', Bel'') \leq 1$,
(MA5) : $Conf(Bel', Bel'') = 0$ iff $m' \subseteq m$" or $m'' \subseteq m'$ [5].

(P3) : Extreme values: $Conf(Bel_1, Bel_2) = 0$ iff $\bigcap_{X\in\mathcal{F}_1\cup\mathcal{F}_2} \neq \emptyset$ iff
iff $\sum_{X\cap Y=\emptyset} m_1(X)m_2(Y) = 0$ iff $\mathcal{P}_{m_1} \cap \mathcal{P}_{m_2} \neq \emptyset$,
where $\mathcal{P}_m = \{Prob \mid Bel(X) \leq Prob(A),\ \forall X \subseteq \Omega\}$;
$Conf(Bel_1, Bel_2) = 1$ iff $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$.
(P4) : Symmetry.
(P5) : Imprecision monotonicity.
(P6) : 'Ignorance is Bliss' $\sim Conf(Bel, Bel_{vac}) = 0$.

conflict correspond to their global conflict.

[4] Let us note, that there are different approaches how to define internal conflict of BFs, e.g., using minnimal entropy functional, see [2], or using author's conflicting parts of BFs [12].

[5] A very special case of belief specialization: $m' \subseteq m$" implies that m' is a specialization of m'', but reverse implication does not hold true.

(P7) : Insensitivity to refinement. (a conflict should not be changed if frame of discernment is refined)

A1 corresponds to axioms MA1 and MA4 and it is consistent with properties P3, P5, P6. A2 corresponds to MA3 and to property P4. A3 corresponds to axiom MA2, this is not assumed by D & B, on the other hand it inconsistent with strong property P3. A4 follows rather strong MA5; A4 corresponds to P6.

4 Former Approaches to Conflict between BFs

To compare the new consonant approach to conflict, let us briefly introduce the former approaches.

4.1 Three Approaches from IPMU 2010

Unfortunately, there are not yet any precise formulas[6], but only bounding inequalities for *combinational conflicts*: $\frac{1}{2}TotC(m,m)) \leq IntC(m) \leq TotC(m,m)$, $TotC(m_1,m_2) - (IntC(m_1) + IntC(m_2)) \leq C(m_1,m_2) \leq TotC(m_1,m_2)$.

Internal plausibility conflict of BF Bel is defined as $Pl\text{-}IntC(Bel) = 1 - max_{\omega\in\Omega} Pl(\{\omega\})$, where Pl is the plausibility equivalent to Bel.

Plausibility conflict between BFs Bel_1, Bel_2 is defined by the formula $Pl\text{-}C(Bel_1, Bel_2) = min(\sum_{\omega\in\Omega_{PlC}(Bel_1,Bel_2)} \frac{1}{2}|Pl\_P(Bel_1) - Pl\_P(Bel_2)(\omega)|, (m_1 \textcircled{\cap} m_2)(\emptyset))$,where $\Omega_{PlC}(Bel_1, Bel_2)$ is the set of elements $\omega \in \Omega$ with conflicting Pl_P masses [11, 14].

The idea of comparative conflictness/non-conflictness is a specification of bbms to smaller focal elements such that fit to focal elements of the other BF as much as possible. *The comparative conflict between BFs* Bel_1 *and* Bel_2 is defined as the least difference of such more specified bbms derived from the input m_1 and m_2.

4.2 Liu's Degree of Conflict and Pignistic Conflict

The above 3 approaches were compared with Liu's *degree of conflict* cf in [11]; cf is defined as $cf(m_i, m_j) = (m_\oplus(\emptyset), difBetP_{m_i}^{m_j})$ in [22], where $m_\oplus(\emptyset)$ should be rather $m_{\textcircled{\cap}}(\emptyset)$ (more precisely $(m_i \textcircled{\cap} m_j)(\emptyset)$) in fact, $difBetP_{m_i}^{m_j}$ is defined as $difBetP_{m_i}^{m_j} = max_{A\subseteq\Omega}(|BetP_{m_i}(A) - BetP_{m_j}(A)|)$. It holds: $difBetP_{m_i}^{m_j} = Diff(BetP_{m_i}, BetP_{m_j}) = \frac{1}{2}\sum_{\omega\in\Omega} |BetP_{m_i}(\{\omega\}) - BetP_{m_j}(\{\omega\})|$ [13].

Pignistic conflict is an alternative of the plausibility conflict [14], where pignistic probability $BetP$ is used instead of normalised plausibility of singletons.

4.3 Conflict Based on Non-Conflicting Parts

For the recent measure of conflict $ncp\text{-}Conf$ [15] is based on Daniel's ideas from [11] and namely from [12]. When analysing properties of approaches from [11] using

[6] Let us recall that notion 'total conflict' $TotC$ is used for global conjunctive conflict $GlcC$ [11].

Hájek-Valdés algebraic approach [19, 20], hypothesis[7] of decomposition of a BF into its conflicting and non-conflicting parts was formulated in [12]; and existence of unique non-conflicting part Bel_0 of any BF Bel was proven there:

Theorem 1. *Let $h(Bel) = Bel \oplus U_n$, where U_n is the uniform distribution on sigletons, i.e., $U_n(\{\omega_i\}) = \frac{1}{n}$. For any BF Bel defined on Ω_n there exists unique consonant BF Bel_0 such that, $h(Bel_0 \oplus Bel_S) = h(Bel)$ for any BF Bel_S such that $Bel_S \oplus U_n = U_n$.*

Definition 1. *Let Bel', Bel'' be two belief functions on n-element frame of discernment $\Omega_n = \{\omega_1, \omega_2, ..., \omega_n\}$. Let Bel'_0 and Bel''_0 be their non-conflicting parts and m'_0, m''_0 the related bbas. We define* conflict between BFs Bel' and Bel'' *as ncp-Conf(Bel',Bel'') $= m_{Bel'_0 \circledcirc Bel''_0}(\emptyset) = (m'_0 \circledcirc m''_0)(\emptyset)$. Where $\circledcirc$ is non-normalised Dempster's (conjunctive) rule of BFs combination.*

For algorithm of computation of *ncp-Conf*(Bel',Bel'') see [15].

5 Consonant Conflicts between BFs

Probabilistic approximations of belief functions were used in several previous approaches, e.g. pignistic probability in Liu's degree of conflict cf and in pignistic conflict $BetP$-C, and normalized plausibility of singletons in plausibility conflict $Pl\!\!\perp P$-C.

Making a probabilistic approximation has two disadvantages in general: the approximation adds some additional information and internal conflict of a BF is increased. As we do not know how internal conflicts of individual BFs participates in global conflict of these BFs, a probabilistic approximation brings also an unspecified contribution to the conflict "between" which is defined using the transformation.

Our present idea is to use consonant approximations $cAppox(Bel)$ instead of the probabilistic ones. Theoretically, we can use any consonant approximation such that the original BF is its specialization. We use more strict condition: inverse probabilistic transformations, i.e., such that $Transf(Bel) = Transf(cAppox(Bel))$, specially for pignistic transformation $BetT(Bel) = BetP$ and plausibility (i.e. contour) transform $Pl\!\!\perp T(Bel) = Pl\!\!\perp P$. Thus for $BetT(Bel) = BetP$ we will use *consonant pignistic inverse* $iBetT(Bel) =_{iBet} Bel =iBet$, given by ${}_{iBet}m$, i.e., consonant inverse of $BetP$: $BetT(iBetT(Bel)) = BetT(Bel) = BetP$ and *consonant inverse contour* iC, i.e., consonant inverse of $Pl\!\!\perp P$: $Pl\!\!\perp T(iCT(Bel)) = Pl\!\!\perp T(iPl\!\!\perp T(Bel)) = Pl\!\!\perp T(Bel) = Pl\!\!\perp P$, see Figure 4.

These approximations have several advantages: they have no internal conflict - the entire conflict is the conflict "in between". No additional information nor internal conflict is added; internally conflicting information is removed. Analogously

[7] This hypothesis have been proven only on BFs on Ω_2 in [12]. In general case, the conflicting part seems to be in a close relationship to internal conflict of the BF.

to the original probabilistic approximations these are also uniquely defined and probabilistic approximation is preserved.

Having consonant approximations of two belief functions, we define a conflict between them. As the internal conflict of consonant approximations is zero, the entire conflict of these approximations is the conflict "in between". Let us adopt the simplest sum of multiples of disjoint focal elements:

Definition 2. *Let Bel_1, Bel_2 be two belief functions on Ω, ${}_{iC}Bel_i = iCT(Bel_i)$ and ${}_{iBet}Bel_i = iBetT(Bel_i)$ be their consonant inverse contour and consonant inverse pignistic approximations given by consonant bbas ${}_{iC}m_i$, ${}_{iBet}m_i$.*
Inverse contour conflict *is defined by formula*

$$iC\text{-}Conf(Bel_1, Bel_2) = \sum_{X \cap Y = \emptyset} {}_{iC}m_1(X) {}_{iC}m_2(Y),$$

where $X, Y \subseteq \Omega$ (i.e., where $X \in \mathcal{F}_{{}_{iC}m_1}$, $Y \in \mathcal{F}_{{}_{iC}m_2}$).
Inverse pignistic conflict *is analogously defined by*

$$iBet\text{-}Conf(Bel_1, Bel_2) = \sum_{X \cap Y = \emptyset} {}_{iBet}m_1(X) {}_{iBet}m_2(Y),$$

where $X, Y \subseteq \Omega$ (i.e., where $X \in \mathcal{F}_{{}_{iBet}m_1}$, $Y \in \mathcal{F}_{{}_{iBet}m_2}$).

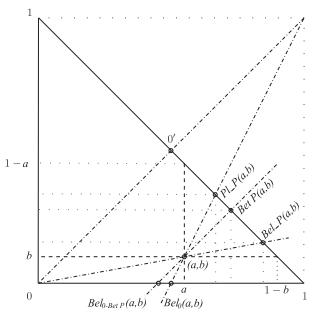


Figure 4: Consonant Approximations on Ω_2

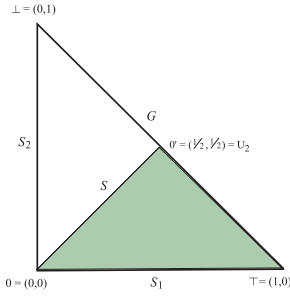


Figure 5: Simplices of mutually non-conflicting Belief functions on Ω_2.

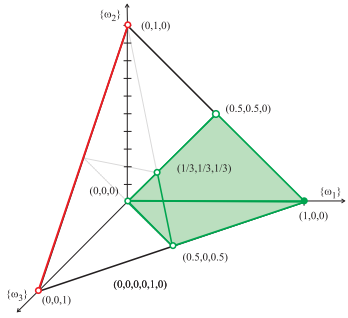


Figure 6: Simplex of quasi Bayesian BFs non-conflicting with $(1, 0, 0)$.

5.1 Basic Properties of iC-$Conf$ and $iBet$-$Conf$

We may easily verify that Definition 2 satisfy assumptions A1 – A4.

Lemma 1. *The following is equivalent: (i) $iC\text{-}Conf(Bel_1, Bel_2) = 0$*
(ii) $\bigcap_{X \in \mathcal{F}_{{}_{iC}m_i}} X \neq \emptyset$
(iii) $X_0 \cap Y_0 \neq \emptyset$ where $X_0 \in \mathcal{F}_{{}_{iC}m_1}, Y_0 \in \mathcal{F}_{{}_{iC}m_2}$ and $X_0 \subset X$, $Y_0 \subset Y$ for any $X \in \mathcal{F}_{{}_{iC}m_1}, Y \in \mathcal{F}_{{}_{iC}m_2}$
(iv) $\{\omega_M | PlP_1(\omega_M) \geq PlP_1(\omega), \omega \in \Omega\} \cap \{\omega_M | PlP_1(\omega_M) \geq PlP_2(\omega), \omega \in \Omega\} \neq \emptyset$
(v) $\{\omega_M | Pl_1(\omega_M) \geq Pl_1(\omega), \omega \in \Omega\} \cap \{\omega_M | Pl_1(\omega_M) \geq Pl_2(\omega), \omega \in \Omega\} \neq \emptyset$.

Lemma 2. *The following is equivalent: (i)* $iBet\text{-}Conf(Bel_1, Bel_2) = 0$
(ii) $\bigcap_{X \in \mathcal{F}_{iBet m_i}} X \neq \emptyset$
(iii) $X_0 \cap Y_0 \neq \emptyset$ *where* $X_0 \in \mathcal{F}_{iBet m_1}$, $Y_0 \in \mathcal{F}_{iBet m_2}$ *and* $X_0 \subset X$, $Y_0 \subset Y$ *for any* $X \in \mathcal{F}_{iBet m_1}, Y \in \mathcal{F}_{iBet m_2}$
(iv) $\{\omega_M | BetP_1(\omega_M) \geq BetP_1(\omega), \omega \in \Omega\} \cap \{\omega_M | BetP_1(\omega_M) \geq BetP_2(\omega), \omega \in \Omega\} \neq \emptyset$.

Lemma 3. *For any pair of BFs on* Ω_2 *and, generally, for any pair of qBBFs* Bel_1 Bel_2 *on* Ω_n *it holds that*
(i) $iC\text{-}Conf(Bel_1, Bel_2) = 0$ *iff* $iBet\text{-}Conf(Bel_1, Bel_2) = 0$ *iff* $\{\omega \mid m_1(\{\omega\}) = max_i\{m_1(\{\omega_i\})\}\} \cap \{\omega \mid m_2(\{\omega\}) = max_i\{m_2(\{\omega_i\})\}\} \neq \emptyset$, *i.e., if* $(a_1 - b_1)(a_2 - b_2) \geq 0$ *in the case of* $Bel_i = (a_i, b_i)$ *on* Ω_2*;*
(ii) $iC\text{-}Conf(Bel_1, Bel_2) \geq iBet\text{-}Conf(Bel_1, Bel_2)$.

From the last condition of Lemma 3 (i) it follows that the following holds in our graphical presentation: any two BFs on Ω_2 from right hand half of the triangle are mutually non-conflicting (there is no conflict between them, see the green part of the triangle on Figure 5; analogously for any BFs from left hand white part.

Analogously, it holds for qBBFs on Ω_n: any two qBBFs from an n-dimensional subsimplex (1/n of the entire simplex of qBBFs, which is defined by $(0, 0, 0, ..., 0)$, and corresponding segment of BBFs where $m(\{\omega^*\}) \geq m(\{\omega\})$ for a given $\omega^* \in \Omega_n$ i.e., $1/n$ of $(n-1)$-dimensional subsimplex of BBFs including $m_{\{\omega^*\}} : m_{\{\omega^*\}}(\{\omega^*\}) = 1$) are mutually non-conflicting. E.g. on Ω_3 and $m_{\{\omega_1\}}(\{\omega_1\}) = 1$ and segment of BBFs given by $(1, 0, 0), (\frac{1}{2}, \frac{1}{2}, 0), (\frac{1}{2}, 0, \frac{1}{2})$, and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, see green subsimplex on Figure 6. Any BF from green subsimplex is conflicting with any BFs from the white part (the rest) of the simplex, i.e. there is some positive conflict between them. For categorical qBBF $(1, 0, 0)$, the maximal value 1 of conflict appears with any BF from red line between $(0, 0, 1)$ and $(0, 1, 0)$. The green subsimplex of mutually non-conflicting BFs is one of maximal consistent simplices discussed in [6].

Corollary 1. *Any symmetric qBBF* Bel_S *on* Ω_n *is non-conflicting with any other qBBF* Bel*, i.e.,* $iC\text{-}Conf(Bel_S, Bel) = 0 = iBet\text{-}Conf(Bel_S, Bel)$.

The situation is significantly more complicated on a $2^{(n-2)}$-dimensional simplex of general BFs on Ω_n. There is multidimensional structure of BFs instead of 1-dimensional *h-line* (a straight line[8] connecting a BF and related Pl_P); i.e., multidimensional structure of BFs with the same Pl_P. Analogously there is a multidimensional structure of BFs with the same $BetP$ instead of simple 1-dimensional perpendicular in the case of qBBFs. A simplex of BFs non-conflicting with $(1, 0, 0, 0, 0, 0)$ analogous to that of Figure 6 is 6-dimensional for general BFs on Ω_3. Thus we have no simple generalization of Lemma 3 to general BFs. On the other hand we can generalize its Corollary to Lemma 4.

Example 1. Let $m_1 = (1, 0, 0, 0, 0, 0)$; for $m_2 = (\frac{1}{3}, 0, 0, 0, 0, \frac{2}{3})$ we obtain $BetP_2 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0) = U_3$, $_{iBet}m_2 = (0, 0, 0, 0, 0, 0) = 0$ thus it is $iBet$-non-conflicting

[8] Its name comes from homomorhism h of algebraic structure of BFs, which is defined by $h(Bel) = Bel \oplus U_n$, see [8, 20].

with m_1; whereas $P\L P_2 = (\frac{1}{5}, \frac{2}{5}, \frac{2}{5}, 0, 0, 0)$, ${}_{iC}m_2 = (0, 0, 0, 0, 0, \frac{2}{3})$, there is $iC\text{-}Conf$ $(m_1,m_2) = \frac{2}{3} > 0 = iBet\text{-}Conf(m_1,m_2)$. On the other hand, for $m_3 = (0,0,\frac{1}{2},\frac{1}{2},0,0)$ we obtain $P\L P_3 = U_3$ thus ${}_{iC}m_2 = 0$, whereas $BetP_3 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0, 0, 0)$, and ${}_{iBet}m_3 = (0, 0, \frac{1}{4}, 0, 0, 0)$, hence $iC\text{-}Conf(m_1, m_3) = 0 < \frac{1}{4} = iBet\text{-}Conf(m_1, m_3)$.

Lemma 4. *(i) Any general symmetric BF Bel_S on Ω_n is non-conflicting with any other BF Bel, i.e., $iC\text{-}Conf(Bel_S, Bel) = 0 = iBet\text{-}Conf(Bel_S, Bel)$.*
(ii) Any BF Bel_{UPl} such that its $P\L P_{UPl} = U_n$ is iC-non-conflicting with any BF Bel on Ω_n, i.e., for any Bel and any Bel_{UPl} with uniform plausibility it holds that $iC\text{-}Conf(Bel_{UPl}, Bel) = 0$.
(iii) Any BF Bel_{UBet} such that its $BetP_{UBet} = U_n$ is iBet-non-conflicting with any BF Bel on Ω_n, i.e., for any Bel and any Bel_{UBet} with uniform BetP holds $iBet\text{-}Conf(Bel_{UBet}, Bel) = 0$.

Lemma 5. *For any BF Bel on Ω_n, its core $\mathcal{C}_m$, cores $\mathcal{C}_{P\L P}$, $\mathcal{C}_{BetP}$ of related probabilities and cores $\mathcal{C}_{iCm}$, $\mathcal{C}_{iBetPm}$ of their consonant approximations hold that $\mathcal{C}_m = \mathcal{C}_{P\L P} = \mathcal{C}_{iCm} = \mathcal{C}_{BetP} = \mathcal{C}_{iBetm}$.*

Lemma 6. *For any pair of BFs Bel_1, Bel_2 on Ω_n and their cores $\mathcal{C}_{m_1}$, $\mathcal{C}_{m_2}$ it holds that (i) $iC\text{-}Conf(Bel_1, Bel_2) = 1$ iff $\mathcal{C}_{m_1} \cap \mathcal{C}_{m_2} = \emptyset$,*
(ii) $iBet\text{-}Conf(Bel_1, Bel_2) = 1$ iff $\mathcal{C}_{m_1} \cap \mathcal{C}_{m_2} = \emptyset$.

5.2 An Equivalence of Consonant $iC\text{-}Conf$ to Conflict between BFs Based on their Non-Conflicting Parts $ncp\text{-}Conf$

Lemma 7. *Consonant inverse contour conflict $iC\text{-}Conf$ is equivalent to conflict between belief functions based on their non-conflicting parts $Conf$, i.e., for any pair of BFs Bel', Bel'' on Ω_n it holds that $iC\text{-}Conf(Bel', Bel'') = ncp\text{-}Conf(Bel', Bel'')$.*

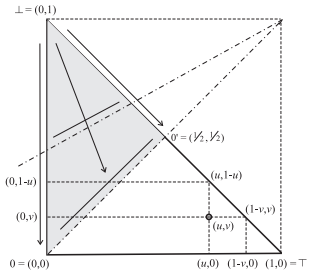


Figure 7: Inverse contour conflict between fixed BF (u, v) and general BF (a, b) on Ω_2; $iC\text{-}Conf((u, v), (a, b))$ decreases in direction of arrows and it is constant along lines without arrows.

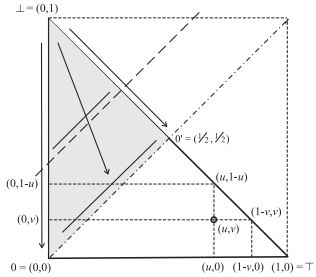


Figure 8: Inverse pignistic conflict between fixed BF (u, v) and general BF (a, b) on Ω_2; $iBet\text{-}Conf((u, v), (a, b))$ decreases in direction of arrows and it is constant along lines without arrows.

Unfortunately, we have found general counterexamples against Theorem 2 from [15], thus the following holds only for qBBFs, it does not hold in general.

Theorem 2. *(i) Let Bel_1 and Bel_2 be arbitrary quasi Bayesian BFs on general finite frame of discernment Ω_n given by bbas m' and m''. For both conflicts iC-$Conf$ and $iBet$-$Conf$ between Bel_1 and Bel_2 it holds that*

$$Conf(Bel_1, Bel_2) \leq \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y).$$

(ii) Equality $Conf(Bel_1,Bel_2) = \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y)$ holds iff both BFs Bel_1 and Bel_2 are consonant.

This statement does not hold for general BFs. see Example 2 [17].

Example 2. *(Counter-example against Theorem 2 from Belief'14 [15] on Ω_3)* Let us suppose Ω_3, $m_1(\{\omega_1, \omega_2\}) = 0.7$, $m_1(\{\omega_1, \omega_3\}) = 0.3$, and $m_2(\{\omega_2, \omega_3\}) = 1.0$. There is $Pl_1 = (1.0, 0.7, 0.3, ...)$, $iC_1 = (0.3, 0, 0, 0.4, 0, 0)$, $Pl_2 = (0, 1.0, 1.0, ...)$, $iC_2 = (0, 0, 0, 0, 0, 1.0)$, thus $iC\text{-}Conf(m_1, m_2) = 0.3 \cdot 1.0 = 0.3$; analogously $BetP_1 = (0.5, 0.35, 0.15)$, $iBet_1 = (0.15, 0, 0, 0.4, 0, 0)$, $BetP_2 = (0, 0.5, 0.5)$, $iBet_2 = (0, 0, 0, 0, 0, 1.0)$, thus $iBet\text{-}Conf(m_1, m_2) = 0.15$.
Nevertheless $\sum_{X \cap Y = \emptyset} m_1(X) m_2(Y) = 0 < 0.15 = iBet\text{-}Conf(m_1, m_2) < 0.30 = iC\text{-}Conf(m_1, m_2)$.

5.3 Relationships to Axiomatic Approaches

We have already seen that Martins axioms MA1 – MA4 are satisfied, due to their correspondence with our satisfied assumptions A1 – A3. MA5 is not satisfied as it is too strong due to Martin's strong definition of bba inclusion, nevertheless our assumption A4 is a consequence of MA5. Martin explicitly does not assume triangle inequality $Conf(Bel',Bel''') \leq Conf(Bel',Bel'') + Conf(Bel'',Bel''')$. Both consonant conflicts are the cases, where triangle inequality does not hold true, see Ex. 3.

Example 3. Let Bel'=(0.4,0.1,0.1,0.2,0,0.1;0.1), Pl'=($\frac{7}{15}$,$\frac{5}{15}$,$\frac{3}{15}$), Bel''=(0.3,0.2,0.1, 0.1,0,0.1;0.2), Pl''=($\frac{6}{16}$,$\frac{6}{16}$,$\frac{4}{16}$), Bel'''=(0.1,0.2,0.3,0.1,0,0.2;0.1), Pl'''=($\frac{3}{15}$,$\frac{6}{15}$,$\frac{6}{15}$), $Bel'_0 = (\frac{2}{7}, 0, 0, \frac{2}{7}, 0, 0; \frac{3}{7})$, $Bel''_0 = (0, 0, 0, \frac{2}{6}, 0, 0; \frac{4}{6})$, $Bel'''_0 = (0, 0, 0, 0, 0, \frac{3}{6}, 0, 0; \frac{3}{6})$, $iC\text{-}Conf(Bel', Bel''') = \frac{1}{7} \not\leq 0 + 0 = iC\text{-}Conf(Bel', Bel'') + iC\text{-}Conf(Bel'', Bel''')$.

P3: We have equivalence only for maximal value; the strongest minimal value condition (i) implies (consonant) non-conflictness in general, medium condition (ii) does it only for quasi Bayesian BFs, and the weakest condition (iii) does not imply non-conflictness at all; either one of reverse implications does not hold true (either for qBBFs). We have validity of P4 and and P6, see our assumptions A2 and A4 above. P5 does not hold either for one of consonant conflicts as specialization of bbms can change order of plausibility and $BetP$ values, thus also focal elements of consonant approximations. P7 is most interesting of the properties, it distinguishes consonant conflicts: it holds for $iC\text{-}Conf$ whereas does not hold for $iBet\text{-}Conf$, due to that plausibility and iC approximations are consistent with refinement of the frame of discernment, but $BetP$ and $iBet$ do not.

Theorem 3. *Let Bel_1, Bel_2 be any BFs given by m_1, m_2 on general Ω_n. For both consonant conflicts iC-Conf and iBet-Conf between Bel_1 and Bel_2 it holds that*
(i) if $\bigcap_{X\in\mathcal{F}_1\cup\mathcal{F}_2}=\emptyset$ then $Conf(Bel_1,Bel_2)=0$,
(ii) if both Bel_1 and Bel_2 are quasi Bayesian and $\sum_{X\cap Y=\emptyset} m_1(X)m_2(Y)=0$ then $Conf(Bel_1,Bel_2)=0$,
(iii) $Conf(Bel_1,Bel_2)=1$ iff $\mathcal{C}_1\cap\mathcal{C}_2=\emptyset$.

5.4 A Comparison with Previous Approaches to Conflict

5.4.1 Combinational Conflict

We suppose $TotC(m_1,m_2)-(IntC(m_1)+IntC(m_2))\leq C(m_1,m_2)\leq TotC(m_1,m_2)$ for combinational conflict. On the other hand, $0\leq Conf(Bel_1,Bel_2)\not\leq \sum_{X\cap Y=\emptyset} m_1(X)m_2(Y) = TotC(Bel_1,Bel_2)$ for both *iC-Conf* and *iBet-Conf*. Thus both the consonant conflicts are not compatible with the formulation of combinational conflict [11].

5.4.2 Plausibility Conflict *Pl-C*

More interesting is a comparison with the most elaborated and precisely defined plausibility conflict. We can observe that: (*Conf* stands for *iC-Conf* or *iBet-Conf*)

Lemma 8. *For any couple of belief functions (a,b), (c,d) on 2-element frame of discernment it holds that: (i) $Conf((a,b),(c,d))=0$ iff Pl-$C((a,b),(c,d))=0$,*
(ii) $Conf((a,b),(c,d))\leq$ Pl-$C((a,b),(c,d))$.
For any couple of belief functions Bel', Bel'' on general finite frame of discernment Ω_n it holds that: (iii) iC-Conf$(Bel',Bel'')=0$ iff Pl-$C_{sm}(Bel',Bel'')=0$.
(iv) If Pl-$C_0(Bel',Bel'')=0$ then also iC-Conf$(Bel',Bel'')=0$ (in general; but not for final Pl-$C(Bel',Bel'')$).
(v), (vi) For qBBFs (iii) + (iv) hold also for iBet-Conf.

Thus, the nature of $Conf((a,b),(c,d))$ is very close to that of Pl-$C((a,b),(c,d))$. $Conf((a,b),(c,d))$ is simpler as its conflictness/non-conflictness simply comes from $\sum_{X\cap Y=\emptyset} {}_{iC}m_1(X)_{iC}m_2(Y)=\frac{|a-b|}{1-min(a,b)}\cdot\frac{|d-c|}{1-min(c,d)}$, $\sum_{X\cap Y=\emptyset} {}_{iBet}m_1(X)_{iBet}m_2(Y)=|a-b|\cdot|d-c|$. Hence there is no necessity to check conflictness of all focal elements. For the same nature see also Figures 7 and 8 which fit also to *Pl-C* and *Bet-C*, respectively.

5.4.3 Comparative Conflict *cp-C*

Comparative conflict has a completely different nature. There are mutually comparatively conflicting couples of BFs with same *max Pl* or *max BetP* elements of Ω_n. Thus they are non-conflicting according to *Conf*, (e.g., $Bel'=(0.5,0.3,0,0,0,0.1)$, $Bel''=(0.7,0,0,0,0.1,0.1)$.) On the other hand, there are comparatively non-conflicting BFs, which prefer different ω_i's - they are conflicting according to *Conf* (e.g. $Bel'=(0.4,0.2,0.1,0.1,0,0)$, $Bel''=(0.3,0.4,0.1,0.1,0,0)$). *cp-C* has some relationship to property P5, which should be investigated in future.

5.4.4 Liu's Measure of Conflict cf

Any couple of BFs that is mutually non-conflicting according to cf (Section 4.2) is (under some conditions) also mutually non-conflicting according to consonant conflicts (the reverse does not hold true); and $Conf$ is less or equal to cf. For behaviour of values of both the components of $cf = (m_{\textcircled{\cap}}(\emptyset), difBetP)$ of a fixed (u, v) with a general (a, b) on Ω_2 see Figures 9, 10; values of both the components decrease in direction of arrows, they are constant along lines without arrow.

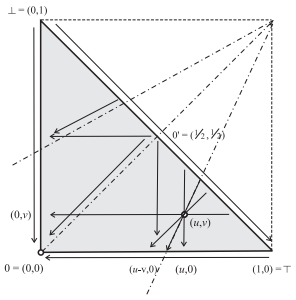


Figure 9: $m_{\textcircled{\cap}}(\emptyset)$ component of cf between fixed (u, v) and general (a, b).

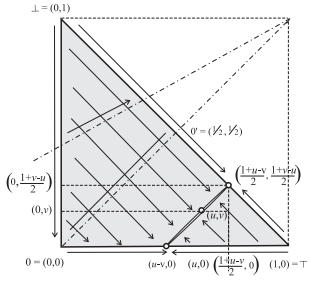


Figure 10: $difBetP$ between fixed BF (u, v) and general BF (a, b) on Ω_2.

Lemma 9. *(i) For any couple of belief functions Bel', Bel'' on n-element frame of discernment it holds that, if $cf(Bel', Bel'') = ((m' \textcircled{\cap} m'')(\emptyset), difBet_{Bel'}^{Bel''}) = (0, 0)$ then also $iBet\text{-}Conf(Bel', Bel'') = 0$,*
(ii) The above holds also for any pair of qBBFs Bel', Bel'' and $iC\text{-}Conf(Bel', Bel'')$
(iii) For $Bel' = (a, b)$, $Bel'' = (c, d)$ on Ω_2 holds further: $Conf((a, b), (c, d)) \leq cf((a, b), (c, d))$; or precisely $Conf \leq ((a, b) \textcircled{\cap} (c, d))(\emptyset)$ & $Conf \leq difBet_{(a,b)}^{(c,d)}$.

6 Conclusion

In this study, we introduced a new approach of conflict between belief functions on general finite frame of discernment. Properties of its instances $iC\text{-}Conf$ and $iBet\text{-}Conf$ were analyzed and compared with former approaches. Conflict based on non-conflicting parts of BFs [15] was observed to be equivalent to consonant conflict $iC\text{-}Conf$. Further, satisfaction of Martin's [23] and Destercke-Burger's [18] axioms was studied.

A common elaboration of the theoretic principles of the presented results with those from [18] and [23] is a challenge for a future research. It should include an analysis of positive conflict in situations such that $\sum_{X \cap Y = \emptyset} m_1(X) m_2(X) = 0$.

The presented theoretical results improve general understanding of conflict between belief functions and entire nature of belief functions. Correct understanding of conflicts may, consequently, improve a combination of conflicting belief functions.

Acknowledgements

This work was supported by the Czech Science Foundation (project 16-12010S).

References

[1] R. G. Almond. *Graphical Belief Modeling.* Chapman & Hall, London (1995)

[2] A. G. Bronevich, I. N. Rozenberg. The choice of generalized Dempster-Shafer rules for aggregating belief functions. *IJAR* **56**, 122–136 (2015)

[3] T. Burger. Geometric Interpretations of Conflict: a Viewpoint. In: F. Cuzzolin (eds.) *BELIEF 2014.* LNCS (LNAI) vol. 8764, pp. 412–421. Springer (2014)

[4] M. E. G. V. Cattaneo. Combining Belief Functions Issued from Dependent Sources. ISiPTA'03, pp. 133–147. Carleton Scientific (2003)

[5] B. R. Cobb, P. P. Shenoy. A Comparison of Methods for Transforming Belief Functions Models to Probability Models. In: T. D. Nielsen, N. L. Zhang (eds.), *ECSQARU 2003.* LNAI 2711, 255–66. Springer (2003)

[6] F. Cuzzolin. On consistent approximations of belief functions in the mass space. In: W. Liu (ed.), *ECSQARU 2011.* LNAI 6717: 287–298, Springer-Verlag, Berlin Heidelberg, 2011.

[7] F. Cuzzolin. Lp consonant approximation of belief functions in the mass space. F. Coolen, et al. (eds.) *ISIPTA'11; Proc. 7th ISIPTA*, pp. 139–148 (2011)

[8] M. Daniel. Algebraic structures related to Dempster-Shafer theory. In: Bouchon-Meunier, B., Yager, R. R., Zadeh, L. A. (eds.) *Advances in Intelligent Computing - IPMU'94.* LNCS 945, 51–61, Springer, Heidelberg (1995)

[9] M. Daniel. Distribution of Contradictive Belief Masses in Combination of Belief Functions. In: B. Bouchon-Meunier, et al. (eds.) *Information, Uncertainty and Fusion*, pp. 431–446. Kluwer Academic Publishers, Boston (2000)

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

[11] M. Daniel. Conflicts within and between Belief Functions. In: E. Hüllermeier, et al. (eds.) *IPMU 2010.* LNAI 6178, 696–705, Springer, Heidelberg (2010)

[12] M. Daniel. Non-conflicting and Conflicting Parts of Belief Functions. F. Coolen, et al. (eds.) *ISIPTA'11; Proc. 7th ISIPTA*, pp. 149–158. SIPTA (2011)

[13] M. Daniel. Properties of Plausibility Conflict of Belief Functions. In: L. Rutkowski, et al. (eds.) *ICAISC 2013, Part I.* LNCS, vol. 7894, pp. 235–246. Springer, Heidelberg (2013)

[14] M. Daniel. Belief Functions: a Revision of Plausibility Conflict and Pignistic Conflict. In: W. Liu, V. S. Subrahmanian, J. Wijsen (eds.) *SUM 2013.* LNCS (LNAI) vol. 8078, pp. 190–203. Springer (2013)

[15] M. Daniel. Conflict between Belief Functions: a New Measure Based on their Non-Conflicting Parts. In: F. Cuzzolin (eds.) *BELIEF 2014*. LNCS (LNAI) vol. 8764, pp. 321–330. Springer, Heidelberg (2014)

[16] M. Daniel. A Comparison of Plausibility Conflict and of Degree of Conflict Based on Amount of Belief Functions. In: *ECSQARU 2015*. LNCS (LNAI) 9161, 440–450. Springer, Heidelberg (2015)

[17] M. Daniel, V. Kratochvíl. Hidden Auto-Conflict in the Theory of Belief Functions In: V. Novák, M. Inuiguchi, M. Štěpnička (eds.) *CJS 2017*. 34–45, IRAFM, University of Ostrava (2017)

[18] S. Destercke and T. Burger. Toward an axiomatic definition of conflict between belief functions. *IEEE Transactions on Cybernetics* **43** (2), 585–596 (2013)

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

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

[21] É. Lefèvre, Z. Elouedi, D. Mercier: Towards an alarm for opposition conflict in a conjunctive combination of belief functions. In: W. Liu (ed.), *ECSQARU 2011*. LNAI 6717: 314–325. Springer, Heidelberg (2011)

[22] W. Liu. Analysing the degree of conflict among belief functions. Artificial Intelligence **170**, 909–924 (2006)

[23] A. Martin. About Conflict in the Theory of Belief Functions. In: T. Denœux, M. H. Masson (eds.) *Belief Functions: Theory and Applications*, AISC 164, pp 161–168. Springer, Berlin Heidelberg (2012)

[24] A. Roquel, S. Le Hégarat-Mascle, I. Bloch, and B. Vincke. Decomposition of conflict as a distribution on hypotheses in the framework on belief functions. *International Journal of Approximate Reasoning* **55** (5), 1129–1146 (2014)

[25] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N. J. (1976)

[26] L. S. Shapley. A Value for n-Person Games. Annals of Mathematics Studies **28**, 307–317, Princeton University Press, Princeton (1953)

[27] P. Smets. The combination of evidence in the transferable belief model. IEEE-Pattern analysis and Machine Intelligence **12**, 447–458 (1990)

[28] P. Smets. Analyzing the combination of conflicting belief functions. *Information Fusion* **8**, 387–412 (2007)

Ranking Alternatives with Tolerance and Reduction in the Setting of Interval AHP*

Masahiro Inuiguchi
Department of Systems Innovation
Graduate School of Engineering Science, Osaka University
e-mail: inuiguti@sys.es.osaka-u.ac.jp

Abstract

Several approaches to the refinement of the dominance relation between alternatives are proposed in the setting of interval AHP in this paper. The approaches are divided into two groups: one uses the tolerance of utility difference and the other uses the reduction of interval priority weights. It is shown that refined dominance relations are obtained relatively easily by solving linear programming problems.

1 Introduction

By the conventional Analytic Hierarchy Process (AHP), alternatives are ranked simply by priority weights estimated from pairwise comparison matrices (PCMs) under multiple criteria [3]. In estimating priority weights, only the pairwise comparison matrices whose consistency degrees are in the acceptable level are treated. Once a priority weight vectors are estimated, the inconsistencies in given pairwise comparison matrices are discarded. From the viewpoint that the decision maker may have vague evaluations, an approach to estimating priority weights by intervals was proposed by Sugihara and Tanaka [5]. Because the intervals estimated by their proposed method do not reflect well the vagueness of the decision maker's evaluations, improved estimation methods have been proposed (see [2]).

Estimating priority weights by intervals is advantageous in making a robust and safe evaluation considering the ambiguity inherent in given PCMs. However, as priority weights are specified only by intervals in these methods, we cannot always rank alternatives clearly. Because the dominance relation between alternatives becomes only a preorder, we cannot judge surely whether an alternative dominates the other for some pairs of alternatives.

*This work was partially supported by JSPS KAKENHI Grant Number 17K18952.

In this paper, we propose several approaches to refining the dominance relation. To rank alternatives clearly under interval priority weights, we introduce two concepts: tolerance in utility difference and reduction of intervals. Tolerance in utility difference assumes that small utility difference is approved. Reduction of intervals assumes that trimming small portion of interval priority weights is accepted. Several conceivable approaches to refining the dominance relation based on those two concepts are proposed. We show that the refined dominance relation can be obtained by solving linear programming problems.

This paper is organized as follows. In the next section, we review the interval AHP and describe a few methods for estimating an interval priority weight vector from a given pairwise comparison matrix. The dominance relation between alternatives is reviewed. In Section 3, approaches to refining the dominance relation are proposed and exemplified. Some concluding remarks are given in Section 4.

2 Interval AHP

We briefly introduce the interval AHP [2, 5] and describe the problem setting of this paper. For the sake of simplicity, we define $N = \{1, 2, \ldots, n\}$ and $N \backslash j = N \backslash \{j\} = \{1, 2, \ldots, j-1, j+1, \ldots, n\}$ for $j \in N$.

As in the conventional AHP [3, 4], we try to estimate the priority weights from a given pairwise comparison matrix A, i.e.,

$$A = \begin{bmatrix} 1 & \cdots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{n1} & \cdots & 1 \end{bmatrix}, \tag{1}$$

where we assume the reciprocity, i.e., $a_{ij} = 1/a_{ji}$, $i, j \in N$. Because the $(i,j)^{\text{th}}$ component a_{ij} of A shows the relative importance of the i^{th} item over the j^{th} item. Theoretically, we have $a_{ij} = w_i/w_j$, $i, j \in N$ for priority weights w_i and w_j of i^{th} and j^{th} items. However, because of the vagueness of human judgement, we assume only $a_{ij} \approx w_i/w_j$, $i, j \in N$, where $\approx$ stands for "approximately equals to". Then, in the conventional AHP, w_i, $i \in N$ are estimated so as to minimize the errors in A.

In the interval AHP [5], we assume that the decision maker may have a vague priority weight vector whose range can be expressed by an interval priority weight vector $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)^{\mathrm{T}}$ rather than a crisp priority weight vector $\boldsymbol{w}$, where $W_i = [w_i^{\mathrm{L}}, w_i^{\mathrm{R}}]$, $i \in N$ and $w_i^{\mathrm{L}} \leq w_i^{\mathrm{R}}$, $i \in N$. The inconsistency is assumed to be caused by this vagueness in evaluation of priority weights. Accordingly, we assume that a_{ij} is obtained as w_i/w_j with randomly chosen $w_i \in W_i$ and $w_j \in W_j$. Therefore, $\boldsymbol{W}$ should satisfy $a_{ij} \in [w_i^{\mathrm{L}}/w_j^{\mathrm{R}}, w_i^{\mathrm{R}}/w_j^{\mathrm{L}}]$, $i, j \in N$, $i < j$. Let $\mathcal{W}(A)$ be the set of all interval weight vectors $\boldsymbol{W}$ satisfying this condition. Moreover, corresponding to the normality condition of $\boldsymbol{w}$ in the conventional AHP, we require the interval weight vector $\boldsymbol{W}$ to satisfy the normality condition, i.e., $\sum_{j \in N \backslash i} w_j^{\mathrm{R}} +$

$w_i^{\mathrm{L}} \geq 1,\ i \in N$ and $\sum_{j \in N \setminus i} w_j^{\mathrm{L}} + w_i^{\mathrm{R}} \leq 1,\ i \in N$. This condition ensures that for any $w_i^{\circ} \in W_i$, there exist $w_j \in W_j,\ j \in N \setminus i$ such that $\sum_{j \in N \setminus i} w_j + w_i^{\circ} = 1$. Let $\mathcal{W}^{\mathrm{N}}$ be the set of all interval weight vectors $\boldsymbol{W}$ satisfying the normality condition.

In the conventional interval AHP [5], interval priority weights $W_i,\ i \in N$ are estimated by solving the following linear programming problem:

$$\underset{\boldsymbol{W}}{\text{minimize}}\{d(\boldsymbol{W}) \mid \boldsymbol{W} \in \mathcal{W}(A) \cap \mathcal{W}^{\mathrm{N}},\ \epsilon \leq w_i^{\mathrm{L}} \leq w_i^{\mathrm{R}},\ i \in N\}, \tag{2}$$

where ϵ is a sufficiently small positive number and $d : \mathcal{W}^{\mathrm{N}} \to [0, +\infty)$ is defined by

$$d(\boldsymbol{W}) = \sum_{i \in N} (w_i^{\mathrm{R}} - w_i^{\mathrm{L}}). \tag{3}$$

$d(\boldsymbol{W})$ shows the sum of widths of interval priority weights $W_i,\ i \in N$ and it has been considered that the smaller $d(\boldsymbol{W})$ the better estimation. Let $\hat{d}$ be the optimal value to Problem (2).

It is shown that the estimated interval priority weights by (2) do not express well the vagueness of decision maker's evaluation. Therefore, several estimation methods [2] improving the quality of estimated intervals have been proposed. Among them, we consider the maximizing minimum range method which estimates the interval priority weights by the following procedure.

⟨1⟩ Solve the following linear programming problem for each $k \in N$:

$$\underset{\boldsymbol{W}}{\text{minimize}}\{d_{\bar{k}}(\boldsymbol{W}) \mid \boldsymbol{W} \in \mathcal{W}(A) \cap \mathcal{W}^{\mathrm{N}},\ \epsilon \leq w_i^{\mathrm{L}} \leq w_i^{\mathrm{R}},\ i \in N\}, \tag{4}$$

where $d_k : \mathcal{W}^{\mathrm{N}} \to [0, +\infty)$ is defined by

$$d_k(\boldsymbol{W}) = \sum_{i \in N \setminus k} (w_i^{\mathrm{R}} - w_i^{\mathrm{L}}). \tag{5}$$

Let $\hat{d}_{\bar{k}}$ be the optimal value to Problem (4).

⟨2⟩ Solve the following two linear programming problems for each $k \in N$:

$$\underset{\boldsymbol{W}}{\text{maximize}}\{w_k^{\mathrm{R}} \mid \boldsymbol{W} \in \mathcal{W}(A) \cap \mathcal{W}^{\mathrm{N}},\ d_{\bar{k}}(\boldsymbol{W}) = \hat{d}_{\bar{k}},\ \epsilon \leq w_i^{\mathrm{L}} \leq w_i^{\mathrm{R}},\ i \in N\}, \tag{6}$$

$$\underset{\boldsymbol{W}}{\text{minimize}}\{w_k^{\mathrm{L}} \mid \boldsymbol{W} \in \mathcal{W}(A) \cap \mathcal{W}^{\mathrm{N}},\ d_{\bar{k}}(\boldsymbol{W}) = \hat{d}_{\bar{k}},\ \epsilon \leq w_i^{\mathrm{L}} \leq w_i^{\mathrm{R}},\ i \in N\}. \tag{7}$$

Let $\hat{w}_i^{\mathrm{L}}(k)$ and $\hat{w}_i^{\mathrm{R}}(k),\ i \in N$ be values of w_i^{L} and $w_i^{\mathrm{R}},\ i \in N$, respectively, at the obtained optimal solution.

⟨3⟩ The interval weights $\check{W}_j = [\check{w}_j^{\mathrm{L}}, \check{w}_j^{\mathrm{R}}],\ j \in N$ are estimated by the following equations:

$$\check{w}_j^{\mathrm{R}} = \max\left\{\hat{w}_j^{\mathrm{R}}(k) \mid k \in N\right\}, \quad \check{w}_j^{\mathrm{L}} = \min\left\{\hat{w}_j^{\mathrm{L}}(k) \mid k \in N\right\}. \tag{8}$$

Because we have $([\hat{w}_1^{\mathrm{L}}(k), \hat{w}_1^{\mathrm{R}}(k)], [\hat{w}_2^{\mathrm{L}}(k), \hat{w}_2^{\mathrm{R}}(k)], \ldots, [\hat{w}_n^{\mathrm{L}}(k), \hat{w}_n^{\mathrm{R}}(k)]) \in \mathcal{W}^{\mathrm{N}}$, $k \in N$, we obtain $([\check{w}_1^{\mathrm{L}}, \check{w}_1^{\mathrm{R}}], [\check{w}_2^{\mathrm{L}}, \check{w}_2^{\mathrm{R}}], \ldots, [\check{w}_n^{\mathrm{L}}, \check{w}_n^{\mathrm{R}}]) \in \mathcal{W}^{\mathrm{N}}$.

Once an interval weight vector $\boldsymbol{W}$ is obtained, we define a dominance relation between alternatives under the assumption that utility values $u_i(o_p)$ of alternatives o_p in view of each criterion are given. We use dominance relation defined by

$$o_p \succsim_O o_q \Leftrightarrow \forall \boldsymbol{w} \in \boldsymbol{W},\ \mathbf{e}^{\mathrm{T}}\boldsymbol{w} = 1;\ \sum_{i \in N}(u_i(o_p) - u_i(o_q))w_i \geq 0, \tag{9}$$

where $\mathbf{e} = (1, 1, \ldots, 1) \in \mathbf{R}^n$. $o_p \succsim_O o_q$ implies that o_p certainly dominates o_q. This dominance relation is only a preorder (reflexive and transitive) because of interval weights. From $\succsim_O$, we obtain a strong dominance relation $\succ_O$ by $o_p \succ o_q \Leftrightarrow o_p \succsim_O o_q$ and $o_q \not\succsim_O o_p$. (9) is rewritten as

$$o_p \succsim_O o_q \Leftrightarrow \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) = \min\left\{ \sum_{i \in N}(u_i(o_p) - u_i(o_q))w_i \,\middle|\, \boldsymbol{w} \in \boldsymbol{W},\ \mathbf{e}^{\mathrm{T}}\boldsymbol{w} = 1 \right\} \geq 0. \tag{10}$$

3 Refining the Dominance Relation

As described above, the dominance relation $\succsim_O$ is usually only a preorder because the dominance relation holds only when an alternative is better than the other for all possible priority weight vectors. The dominance relation indicated by $\succsim_O$ is the result of careful consideration. Therefore, $\succsim_O$ is useful in knowing the robust dominance relation. However, because we may neither rank alternatives nor find the best alternative by using $\succsim_O$, $\succsim_O$ is weak in giving some guidance or instruction for good evaluation and decision. In this section, we investigate the ways to provide some guidance for ranking alternatives. To this end, we propose several methods for ranking alternatives in the presence of interval priority weights. Two approaches are conceivable: one is based on the tolerance of utility differences and the other is based on the reduction of interval priority weights.

3.1 Tolerance approach

3.1.1 By the minimum utility difference

The first approach uses minimum utility differences between alternatives. The objective function of the optimization problem appears in (10) shows the minimum utility difference of alternative o_p from alternative o_q. By exchanging o_p and o_q, we obtain the minimum utility difference of alternative o_q from alternative o_p which shows also minus 1 times of the maximum utility difference of alternative o_p from alternative o_q. Therefore, by solving the optimization problem which is a linear programming problem appears in (10) twice, we obtain the range of utility difference of alternative o_p from alternative o_q as $[\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q), \delta^{\mathrm{R}}_{\boldsymbol{W}}(o_p, o_q)]$, where $\delta^{\mathrm{R}}_{\boldsymbol{W}}(o_p, o_q) = -\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_q, o_p)$.

In (10), if the minimum utility difference is non-negative, i.e., $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq 0$, we are sure that o_p is not worse than o_q, i.e., $o_p \succsim_O o_q$. From this definition, we may relax the condition $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq 0$ to $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq -\alpha$, $\alpha > 0$ is a small number. By this way, we define a relaxed dominance relation as follows:

$$\succsim^{\mathrm{L}}_{\sim\alpha} = \{(o_p, o_q) \mid \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq -\alpha\}. \tag{11}$$

Namely, we have $o_p \succsim^{\mathrm{L}}_{\sim\alpha} o_q$ if and only if $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq -\alpha$, where $o_p \succsim^{\mathrm{L}}_{\sim\alpha} o_q$ implies that we are sure that o_p is not very much worse than o_q.

As α increases, $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq -\alpha$ holds for more ordered pairs (o_p, o_q). When α exceeds a certain value, we may obtain $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq -\alpha$ and $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_q, o_p) \geq -\alpha$. Namely, we have $o_p \succsim^{\mathrm{L}}_{\sim\alpha} o_q$ and $o_q \succsim^{\mathrm{L}}_{\sim\alpha} o_p$, i.e., o_p and o_q are indifferent by discarding utility difference α. However, this is not always good if $|\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) - \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_q, o_p)|$ is sufficiently large comparing to $\max(-\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q), -\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_q, o_p))$. To avoid this, we modify the definition of $\succsim^{\mathrm{L}}_{\sim\alpha}$ as

$$\succsim^{\mathrm{L}}_{\sim\alpha} = \{(o_p, o_q) \mid \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq -\alpha \text{ and } \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) > \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_q, o_p)\}. \tag{12}$$

Moreover, although this modification is applied with a sufficient large α, $\succsim^{\mathrm{L}}_{\sim\alpha}$ cannot always satisfy the transitivity. In other words, the transitive closure $\mathrm{Trcl}(\succsim^{\mathrm{L}}_{\sim\alpha})$ includes indifferences among many alternatives. When all values of $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q)$ are different, we modify again $\succsim^{\mathrm{L}}_{\sim\alpha}$ by

$$\begin{aligned}\succsim^{\mathrm{L}}_{\sim\alpha} = \{(o_p, o_q) \mid \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) \geq -\alpha \text{ and } (\forall \zeta < \alpha,\ \forall (o_r, o_s) \in \mathrm{Trcl}(\succsim^{\mathrm{L}}_{\sim\zeta} \cup \{(o_p, o_q)\}),\\ \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_r, o_s) \geq -\alpha \text{ or } (o_s, o_r) \notin \mathrm{Trcl}(\succsim^{\mathrm{L}}_{\sim\zeta}))\},\end{aligned} \tag{13}$$

where $\mathrm{Trcl}(\cdot)$ stands for the transitive closure. When k pairs (o_p, o_q) take a same value $\bar{\alpha}$, we introduce some ranking among the k pairs and modify $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q)$ with $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) + (l-1)\epsilon$, where l shows that pair (o_p, o_q) is ranked as the l-th among the k pairs and ϵ is a very small number. As an example of such an extra ranking, we may order the k pairs in increasing order of $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_q, o_p)$. As the result, we obtain a weak order $\mathrm{Trcl}(\succsim^{\mathrm{L}}_{\sim\alpha})$ with a sufficient large number α. We select basically the minimum α such that $\mathrm{Trcl}(\succsim^{\mathrm{L}}_{\sim\alpha})$ with $\succsim^{\mathrm{L}}_{\sim\alpha}$ of (13) becomes a weak order.

Example 1. Consider a multiple criteria decision making problem with five criteria $C_1, \ldots, C_5$ and five alternatives $o_1, \ldots, o_5$. We assume the evaluations in view of each criterion is given as in Table 1. Let U be the matrix shown in Table 1. To obtain priority weights of criteria, we asked the decision maker to give a pairwise comparison matrix (PCM). The obtained PCM is shown in Table 2. The consistency index (C.I.) of the PCM is obtained as 0.05209. Because C.I. is smaller than 0.1, we may regard the given PCM is meaningful (see [3]). Applying the maximum eigenvalue method and the geometric mean method used often in the conventional AHP, we obtain the following priority weight vectors, respectively: $\boldsymbol{w}^{\mathrm{E}} = (0.3558, 0.2394, 0.1578, 0.1349, 0.1121)^{\mathrm{T}}$ and $\boldsymbol{w}^{\mathrm{G}} = (0.3468, 0.2424, 0.1599, 0.1392, 0.1117)^{\mathrm{T}}$. The total scores of alternatives are obtained as $U\boldsymbol{w}^{\mathrm{E}} = (0.2192, 0.2087,$

Table 1: Scores of alternatives

	C_1	C_2	C_3	C_4	C_5
o_1	0.25	0.3	0.1	0.15	0.2
o_2	0.2	0.25	0.3	0.1	0.15
o_3	0.15	0.2	0.25	0.3	0.1
o_4	0.1	0.15	0.2	0.25	0.3
o_5	0.3	0.1	0.15	0.2	0.25

Table 2: Pairwise comparison matrix

	C_1	C_2	C_3	C_4	C_5
C_1	1	1	2	2	6
C_2	1	1	1	2	2
C_3	1/2	1	1	1	1
C_4	1/2	1/2	1	1	1
C_5	1/6	1/2	1	1	1

Table 3: $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q)$

	o_1	o_2	o_3	o_4	o_5
o_1	–	−0.018182	−0.031031	0.008553	−0.003947
o_2	−0.01875	–	−0.033333	−0.010197	−0.022697
o_3	−0.0375	−0.01875	–	0.004546	−0.041447
o_4	−0.077083	−0.058333	−0.039583	–	−0.057143
o_5	−0.028947	−0.045833	−0.027083	0.0125	–

$0.1924, 0.1704, 0.2094)^{\mathrm{T}}$ and $U\boldsymbol{w}^{\mathrm{G}} = (0.2186, 0.2086, 0.1934, 0.1713, 0.2080)^{\mathrm{T}}$. Then, we obtain $o_1 \succ^{\mathrm{E}} o_5 \succ^{\mathrm{E}} o_2 \succ^{\mathrm{E}} o_3 \succ^{\mathrm{E}} o_4$ and $o_1 \succ^{\mathrm{G}} o_2 \succ^{\mathrm{G}} o_3 \succ^{\mathrm{G}} o_5 \succ^{\mathrm{E}} o_4$, respectively. We note that the orders are different between the maximum eigenvalue method and the geometric mean method although C.I. is small enough.

Now we apply the interval AHP. Estimating the interval priority weights by the maximizing minimum range method, we obtain $\boldsymbol{W} = ([0.25, 0.4286], [0.1842, 0.3158], [0.125, 0.2727], [0.125, 0.3333], [0.04167, 0.1818])^{\mathrm{T}}$. $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q)$ are obtained as shown in Table 3. Then we obtain only $o_1 \succ_O o_4$, $o_3 \succ_O o_4$ and $o_5 \succ_O o_4$ when $\alpha = 0$. Setting $\alpha = 0.031031$ or larger, we obtain a weak order defined by $\mathrm{Trcl}(\succsim^{\mathrm{L}}_{\alpha})$. When $\alpha = 0.031031$, we have

$$\succsim^{\mathrm{L}}_{\alpha} = \{(o_1, o_4), (o_3, o_4), (o_5, o_4), (o_1, o_5), (o_2, o_4), (o_1, o_2), (o_3, o_2), (o_2, o_5), (o_1, o_3)\}. \tag{14}$$

By the transitive closure of this relation, we obtain $o_1 \mathrel{\hat{\succsim}^{\mathrm{L}}_{\alpha}} o_3 \mathrel{\hat{\succsim}^{\mathrm{L}}_{\alpha}} o_2 \mathrel{\hat{\succsim}^{\mathrm{L}}_{\alpha}} o_5 \mathrel{\hat{\succsim}^{\mathrm{L}}_{\alpha}} o_4$, where $\hat{\succsim}^{\mathrm{L}}_{\alpha} = \mathrm{Trcl}(\succsim^{\mathrm{L}}_{\alpha})$. □

We note that in the obtained refinement of dominance relation we understand $o_p \succsim^{\mathrm{L}}_{\alpha} o_q$ is accepted more easily if $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q)$ is larger.

3.1.2 By the center value of the utility difference

In the previous subsection, we refined the dominance relation by the minimum value of the utility difference. However, the minimum value can be considerably small if the width of the interval of utility difference is large even if the location of the interval is around zero. For example, in Example 1, the location of the utility difference between o_1 and o_2 is around zero because the difference between

$\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_1, o_2)$ and $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_2, o_1)$ is very small. The location can be seen by the center value of the interval. Then, in this subsection, we consider a refinement by using the center values of utility difference intervals. Because the range of the utility difference of alternative o_p from alternative o_q is obtained as $[\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q), \delta^{\mathrm{R}}_{\boldsymbol{W}}(o_p, o_q)]$. Then the center value of the utility difference of alternative o_p from alternative o_q can be obtained by

$$\delta^{\mathrm{C}}_{\boldsymbol{W}}(o_p, o_q) = \frac{1}{2}(\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) + \delta^{\mathrm{R}}_{\boldsymbol{W}}(o_p, o_q)). \tag{15}$$

As we have $\delta^{\mathrm{R}}_{\boldsymbol{W}}(o_p, o_q) = -\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_q, o_p)$, we obtain $\delta^{\mathrm{C}}_{\boldsymbol{W}}(o_p, o_q) = -\delta^{\mathrm{C}}_{\boldsymbol{W}}(o_q, o_p)$. Because of this special relation, each ordered pair (o_p, o_q) such that $\delta^{\mathrm{C}}_{\boldsymbol{W}}(o_p, o_q) \geq 0$ is a candidate of the refined dominance relation $o_p \succsim^{\mathrm{C}} o_q$. We note that $o_p \succsim_O o_q$ is always a candidate of $o_p \succsim^{\mathrm{C}} o_q$. However, unfortunately, the dominance relation composed of the candidates does not always become a weak order. To overcome this inadequacy, we apply the same idea as $\succsim^{\mathrm{L}}_{\alpha}$. Namely, for $\eta \geq 0$, we define

$$\begin{aligned}\succsim^{\mathrm{C}}_{\eta} = \{(o_p, o_q) \mid \delta^{\mathrm{C}}_{\boldsymbol{W}}(o_p, o_q) \geq -\eta \text{ and } (\forall \zeta < \eta,\ \forall (o_r, o_s) \in \mathrm{Trcl}(\succsim^{\mathrm{C}}_{\zeta} \cup \{(o_p, o_q)\}),\\ \delta^{\mathrm{C}}_{\boldsymbol{W}}(o_r, o_s) \geq -\eta \text{ or } (o_s, o_r) \notin \mathrm{Trcl}(\succsim^{\mathrm{C}}_{\zeta}))\}.\end{aligned} \tag{16}$$

We select η by the minimum value such that $\mathrm{Trcl}(\succsim^{\mathrm{C}}_{\eta})$ becomes a weak order.

Applying the approach of $\succsim^{\mathrm{C}}_{\eta}$ in Example 1, we obtain the same refined weak order $o_1 \hat{\succsim}^{\mathrm{C}}_{\eta} o_3 \hat{\succsim}^{\mathrm{C}}_{\eta} o_2 \hat{\succsim}^{\mathrm{C}}_{\eta} o_5 \hat{\succsim}^{\mathrm{C}}_{\eta} o_4$ with $\eta = 0.0032345$, where $\hat{\succsim}^{\mathrm{C}}_{\eta} = \mathrm{Trcl}(\succsim^{\mathrm{C}}_{\eta})$. However, $\succsim^{\mathrm{C}}_{\eta}$ is different from $\succsim^{\mathrm{L}}_{\alpha}$ in (14), i.e., we have

$$\succsim^{\mathrm{C}}_{\eta} = \{(o_1, o_4), (o_5, o_4), (o_2, o_4), (o_3, o_4), (o_1, o_5), (o_2, o_5), (o_3, o_2), (o_1, o_3)\}. \tag{17}$$

As is shown in (17), (o_1, o_2) does not appear in $\succsim^{\mathrm{C}}_{\eta}$ of (17) while it appears in $\succsim^{\mathrm{L}}_{\alpha}$ of (14). This exemplifies the case where $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_1, o_2)$ is rather large but $\delta^{\mathrm{R}}_{\boldsymbol{W}}(o_1, o_2) - \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_1, o_2)$ is small.

When we use the center values, a simpler approach is conceivable. It utilizes the average degree of dominance

$$avdd(o_p) = \frac{1}{n-1} \sum_{q \in N \setminus p} \delta^{\mathrm{C}}_{\boldsymbol{W}}(o_p, o_q). \tag{18}$$

The larger the average degree of dominance is, the larger we consider its utility is. Therefore, we may rank alternatives by $avdd(o_p)$. This order is denoted by $\tilde{\succsim}^{\mathrm{C}}$. We note that $\sum_{p \in N} avdd(o_p) = 0$. Therefore, we may regard o_p as a preferable alternative if $avdd(o_p) > 0$.

Applying this approach to Example 1, we obtain $avdd(o_1) = 0.058837$, $avdd(o_2) = 0.028061$, $avdd(o_3) = 0.018940$, $avdd(o_4) = -0.123772$ and $avdd(o_5) = 0.017936$. Then the refined dominance relations is obtained as $o_1 \tilde{\succsim}^{\mathrm{C}} o_2 \tilde{\succsim}^{\mathrm{C}} o_3 \tilde{\succsim}^{\mathrm{C}} o_5 \tilde{\succsim}^{\mathrm{C}} o_4$. We note that $avdd(o_1)$ and $avdd(o_2)$ are sufficiently different although $\delta^{\mathrm{C}}_{\boldsymbol{W}}(o_1, o_2)$ is very small.

Table 4: $prt(o_p, o_q)$

	o_1	o_2	o_3	o_4	o_5
o_1	–	0.507690	0.547198	1	0.880009
o_2	0.492310	–	0.360002	0.851204	0.668802
o_3	0.452802	0.639998	–	1	0.395199
o_4	0	0.148796	0	–	0
o_5	0.119991	0.331198	0.604801	1	–

3.2 By the positive ratio of the interval utility difference

In the approach using the center value of interval utility difference, the width of interval utility difference is not taken care at all. Third approach is to take care of the location and the width of interval utility difference. We consider the positive ratio of the interval utility difference. Namely, we calculate the ratio of positive region to the whole range of possible utility differences, i.e.,

$$prt(o_p, o_q) = \frac{\max(\delta^{\mathrm{R}}_{\boldsymbol{W}}(o_p, o_q), 0) - \max(\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q), 0)}{\delta^{\mathrm{R}}_{\boldsymbol{W}}(o_p, o_q) - \delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q)} \tag{19}$$

We note that we have $prt(o_p, o_q) = 1$ if and only if $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q) > 0$, and we have $prt(o_p, o_q) > 0.5$ if and only if $\delta^{\mathrm{C}}_{\boldsymbol{W}}(o_p, o_q) > 0$.

We apply the same idea as $\succsim^{\mathrm{L}}_{\alpha}$ and $\succsim^{\mathrm{C}}_{\eta}$. Then, for $\rho \geq 0.5$, we define

$$\begin{aligned}\succsim^{\mathrm{P}}_{\rho} = \{(o_p, o_q) \mid\ & prt(o_p, o_q) \geq \rho \text{ and } (\forall \zeta < \rho,\ \forall (o_r, o_s) \in \mathrm{Trcl}(\succsim^{\mathrm{P}}_{\zeta} \cup \{(o_p, o_q)\}),\\ & prt(o_r, o_s) \geq \rho \text{ or } (o_s, o_r) \notin \mathrm{Trcl}(\succsim^{\mathrm{P}}_{\zeta}))\}.\end{aligned} \tag{20}$$

As we decrease ρ, we obtain a weak order $\mathrm{Trcl}(\succsim^{\mathrm{P}}_{\rho})$. We select ρ basically with the minimum value such that $\mathrm{Trcl}(\succsim^{\mathrm{P}}_{\rho})$ becomes a weak order.

Example 2. Consider the same pairwise comparison matrix and normalized interval weight vector $\boldsymbol{W}$ as in Example 1. Based on $\delta^{\mathrm{L}}_{\boldsymbol{W}}(o_p, o_q)$ values shown in Table 3, we obtain $prt(o_p, o_q)$ as shown in Table 4. Applying the approach of $\succsim^{\mathrm{P}}_{\rho}$, from Table 4, we obtain a refined weak order, $o_1 \mathrel{\hat{\succsim}^{\mathrm{P}}_{\rho}} o_3 \mathrel{\hat{\succsim}^{\mathrm{P}}_{\rho}} o_2 \mathrel{\hat{\succsim}^{\mathrm{P}}_{\rho}} o_5 \mathrel{\hat{\succsim}^{\mathrm{P}}_{\rho}} o_4$ with $\rho = 0.547198$ or larger. When $\rho = 0.547198$,

$$\succsim^{\mathrm{P}}_{\rho} = \{(o_1, o_4), (o_3, o_4), (o_5, o_4), (o_1, o_5), (o_2, o_4), (o_2, o_5), (o_3, o_2), (o_1, o_3)\}. \tag{21}$$

Pair (o_1, o_2) appear neither in $\succsim^{\mathrm{P}}_{\rho}$.

3.3 Reduction approach

Several approaches to refining dominance relation $\succsim_O$ based on utility difference have proposed in the previous subsection. As another approach, a method based on

interval weight reduction is conceivable (see [1]). In this subsection, the approach based on the reduction of interval weights is described.

Let $\boldsymbol{V} = (V_1, V_2, \ldots, V_n)^{\mathrm{T}}$ be a reduced interval priority weight vector of a given interval weight vector $\boldsymbol{W}$ such that $w_i^{\mathrm{L}} \leq v_i^{\mathrm{L}} \leq v_i^{\mathrm{R}} \leq w_i^{\mathrm{R}}$, $i \in N$, where $V_i = [v_i^{\mathrm{L}}, v_i^{\mathrm{R}}]$, $i \in N$. For o_p to dominate o_q, the reduced interval weight vector $\boldsymbol{V} \subseteq \boldsymbol{W}$ should satisfy

$$\min \left\{ \sum_{i \in N} (u_i(o_p) - u_i(o_q)) v_i \;\middle|\; v_i^{\mathrm{L}} \leq v_i \leq v_i^{\mathrm{R}},\ i \in N,\ \sum_{i \in N} v_i = 1 \right\} \geq 0. \tag{22}$$

Let us define the following three index sets of N:

$$I^+(o_p, o_q) = \{i \in N \mid u_i(o_p) - u_i(o_q) > 0\}, \tag{23}$$
$$I^-(o_p, o_q) = \{i \in N \mid u_i(o_p) - u_i(o_q) < 0\}, \tag{24}$$
$$I^0(o_p, o_q) = \{i \in N \mid u_i(o_p) - u_i(o_q) = 0\}. \tag{25}$$

v_i^{L} for $i \in I^+(o_p, o_q)$ and v_i^{R} for $i \in I^-(o_p, o_q)$ tend to minimize the objective function of the minimization problem in (22). Indeed, if we drop the constraint $\sum_{i \in N} v_i = 1$ from the minimization problem in (22), v_i^{L} for $i \in I^+(o_p, o_q)$ and v_i^{R} for $i \in I^-(o_p, o_q)$ attain the minimum. From this fact, we take care of the changes of the lower bounds of interval priority weights when $u_i(o_p) > u_i(o_q)$ and the changes of the upper bounds of interval priority weights when $u_i(o_p) < u_i(o_q)$.

Then we define the ambiguity reduction rates of $\boldsymbol{V} \subseteq \boldsymbol{W}$ in the following two ways:

individual ambiguity reduction rate:

$$ir(o_p, o_q) = \min \left(\min_{i \in I^-(o_p, o_q)} \frac{w_i^{\mathrm{R}} - v_i^{\mathrm{R}}}{w_i^{\mathrm{R}} - w_i^{\mathrm{L}}}, \min_{i \in I^+(o_p, o_q)} \frac{v_i^{\mathrm{L}} - w_i^{\mathrm{L}}}{w_i^{\mathrm{R}} - w_i^{\mathrm{L}}} \right),$$

total ambiguity reduction rate:

$$tr(o_p, o_q) = \frac{\displaystyle\sum_{i \in I^-(o_p, o_q)} (w_i^{\mathrm{R}} - v_i^{\mathrm{R}}) + \sum_{i \in I^+(o_p, o_q)} (v_i^{\mathrm{L}} - w_i^{\mathrm{L}})}{\displaystyle\sum_{i \in I^-(o_p, o_q) \cup I^+(o_p, o_q)} (w_i^{\mathrm{R}} - w_i^{\mathrm{L}})}. \tag{26}$$

In this paper, we will find, for each ordered pair (o_p, o_q) of alternatives, the reduced interval priority weight vector $\boldsymbol{V}(o_p, o_q)$ which maximizes an ambiguity reduction rate $rd(o_p, o_q)$ such that $\forall \boldsymbol{V'} \supseteq \boldsymbol{V}(o_p, o_q)$ satisfying $\boldsymbol{V'} \subseteq \boldsymbol{W}$, $\boldsymbol{V'} \in \mathcal{W}^{\mathrm{N}}$ and $\delta_{\boldsymbol{V'}}^{\mathrm{L}} \geq 0$. In other words, we maximize an ambiguity reduction rate $rd(o_p, o_q)$ such that $\exists \boldsymbol{v} = (v_1, \ldots, v_n)^{\mathrm{T}} \in \boldsymbol{V}(o_p, o_q)$ satisfying $\mathbf{e}^{\mathrm{T}} \boldsymbol{v} = 1$ and $\sum_{i \in N} (u_i(o_p) - u_i(o_q)) v_i \leq 0$. For $rd(o_p, o_q)$, we consider $ir(o_p, o_q)$ and $tr(o_p, o_q)$.

The maximum $ir(o_p, o_q)$ and $tr(o_p, o_q)$ as well as their corresponding $\boldsymbol{V}(o_p, o_q)$

can be obtained by solving the following linear programming problems, respectively:

$$\begin{aligned}
\text{maximize} \quad & r, \\
\text{sub. to} \quad & \sum_{i\in N}(u_i(o_p)-u_i(o_q))v_i \le 0, \\
& \sum_{i\in N} v_i = 1,\ v_i^{\mathrm{L}} \le v_i \le v_i^{\mathrm{R}},\ i\in N, \\
& v_i^{\mathrm{L}} - (w_i^{\mathrm{R}}-w_i^{\mathrm{L}})r \ge w_i^{\mathrm{L}},\ v_i^{\mathrm{R}} \le w_i^{\mathrm{R}},\ i\in I^+(o_p,o_q), \\
& v_i^{\mathrm{R}} + (w_i^{\mathrm{R}}-w_i^{\mathrm{L}})r \le w_i^{\mathrm{R}},\ v_i^{\mathrm{L}} \ge w_i^{\mathrm{L}},\ i\in I^-(o_p,o_q), \\
& v_i^{\mathrm{L}} \ge w_i^{\mathrm{L}},\ v_i^{\mathrm{R}} \le w_i^{\mathrm{R}},\ i\in I^0(o_p,o_q), \\
& v_i^{\mathrm{L}} + \sum_{j\in N\setminus j} v_i^{\mathrm{R}} \ge 1,\ v_i^{\mathrm{R}} + \sum_{j\in N\setminus j} v_i^{\mathrm{L}} \le 1,\ i\in N, \\
& v_i^{\mathrm{L}} \ge \epsilon,\ i\in N,\ r\ge 0,
\end{aligned} \tag{27}$$

and

$$\begin{aligned}
\text{maximize} \quad & \sum_{i\in I^+(o_p,o_q)\cup I^-(o_p,o_q)} r_i \Bigg/ \sum_{i\in I^+(o_p,o_q)\cup I^-(o_p,o_q)} (w_i^{\mathrm{R}}-w_i^{\mathrm{L}}), \\
\text{sub. to} \quad & \sum_{i\in N}(u_i(o_p)-u_i(o_q))v_i \le 0, \\
& \sum_{i\in N} v_i = 1,\ v_i^{\mathrm{L}} \le v_i \le v_i^{\mathrm{R}},\ i\in N, \\
& v_i^{\mathrm{L}} - r_i \ge w_i^{\mathrm{L}},\ v_i^{\mathrm{R}} \le w_i^{\mathrm{R}},\ i\in I^+(o_p,o_q), \\
& v_i^{\mathrm{R}} + r_i \le w_i^{\mathrm{R}},\ v_i^{\mathrm{L}} \ge w_i^{\mathrm{L}},\ i\in I^-(o_p,o_q), \\
& v_i^{\mathrm{L}} \ge w_i^{\mathrm{L}},\ v_i^{\mathrm{R}} \le w_i^{\mathrm{R}},\ i\in I^0(o_p,o_q), \\
& v_i^{\mathrm{L}} + \sum_{j\in N\setminus j} v_i^{\mathrm{R}} \ge 1,\ v_i^{\mathrm{R}} + \sum_{j\in N\setminus j} v_i^{\mathrm{L}} \le 1,\ i\in N, \\
& v_i^{\mathrm{L}} \ge \epsilon,\ i\in N,\ r_i\ge 0,\ i\in I^+(o_p,o_q)\cup I^-(o_p,o_q).
\end{aligned} \tag{28}$$

We obtain $ir(o_p,o_q)$ and $tr(o_p,o_q)$ by optimal values of Problems (27) and (28), respectively. For each of those problems, the reduced interval priority weight vector $\boldsymbol{V}(o_p,o_q) = (V_1(o_p,o_q),\ldots,V_n(o_p,o_q))^{\mathrm{T}}$ are obtained by $V_i(o_p,o_q) = [v_i^{\mathrm{L}}, v_i^{\mathrm{R}}]$, $i\in N$ from an optimal solution.

For $ir(o_p,o_q)$ and $ir(o_q,o_p)$, we have $ir(o_p,o_q)+ir(o_q,o_p)\le 1$, and for $tr(o_p,o_q)$ and $tr(o_q,o_p)$, we have $tr(o_p,o_q)+tr(o_q,o_p)\le 1$. These can be proven as follows: we show $tr(o_p,o_q)+tr(o_q,o_p)\le 1$. Let $\boldsymbol{r}^* = (r_1^*,\ldots,r_n^*)^{\mathrm{T}}$, $\boldsymbol{v}^{\mathrm{L}*} = (v_1^{\mathrm{L}*},\ldots,v_n^{\mathrm{L}*})^{\mathrm{T}}$ and $\boldsymbol{v}^{\mathrm{R}*} = (v_1^{\mathrm{R}*},\ldots,v_n^{\mathrm{R}*})^{\mathrm{T}}$ compose an optimal solution to Problem (28). We have $tr(o_p,o_q) = \sum_{i\in I^+(o_p,o_q)\cup I^-(o_p,o_q)} r_i$, $v_i^{\mathrm{L}*} = w_i^{\mathrm{L}*} + r_i$, $i\in I^+(o_p,o_q)$, $v_i^{\mathrm{R}*} = w_i^{\mathrm{R}} + r_i$, $i\in I^-(o_p,o_q)$ and $\forall \boldsymbol{v} = (v_1,\ldots,v_n)^{\mathrm{T}}$ such that $\boldsymbol{v}^{\mathrm{L}} \le \boldsymbol{v} \le \boldsymbol{v}^{\mathrm{R}}$, we have $\sum_{i\in N}(u_i(o_p)-u_i(o_q))v_i \le 0$. From the last property, we obtain

$$tr(o_q,o_p) \le \frac{\displaystyle\sum_{i\in I^+(o_q,o_p)\cup I^-(o_q,o_p)} (w_i^{\mathrm{R}} - v_i^{\mathrm{R}*})}{\displaystyle\sum_{i\in I^+(o_q,o_p)\cup I^-(o_q,o_p)} (w_i^{\mathrm{R}} - w_i^{\mathrm{L}})}. \tag{29}$$

Because $I^+(o_p,o_q)=I^-(o_q,o_p)$ holds. Then we obtain $w_i^{\mathrm{R}}-v_i^{\mathrm{R}*}\le(w_i^{\mathrm{R}}-w_i^{\mathrm{L}})-r_i$, $i\in I^+(o_p,o_q)$ and $v_i^{\mathrm{L}*}-w_i^{\mathrm{L}}\le(w_i^{\mathrm{R}}-w_i^{\mathrm{L}})-r_i$, $i\in I^-(o_p,o_q)$. Therefore, we obtain

$$\frac{\displaystyle\sum_{i\in I^+(o_q,o_p)\cup I^-(o_q,o_p)}(w_i^{\mathrm{R}}-v_i^{\mathrm{R}*})}{\displaystyle\sum_{i\in I^+(o_q,o_p)\cup I^-(o_q,o_p)}(w_i^{\mathrm{R}}-w_i^{\mathrm{L}})}\le\frac{\displaystyle\sum_{i\in I^+(o_q,o_p)\cup I^-(o_q,o_p)}(w_i^{\mathrm{R}}-w_i^{\mathrm{L}})-r_i}{\displaystyle\sum_{i\in I^+(o_q,o_p)\cup I^-(o_q,o_p)}(w_i^{\mathrm{R}}-w_i^{\mathrm{L}})}$$
$$=1-tr(o_p,o_q). \tag{30}$$

From (29) and (30), we conclude $tr(o_q,o_p)+tr(o_p,o_q)\le 1$. The other can be proven in the same way.

The smaller $ir(o_p,o_q)$ and $tr(o_p,o_q)$ are, the more acceptable $o_p\succsim o_q$ is. Then we refine $\succsim_O$ by accepting $o_p\succsim o_q$ with small $ir(o_p,o_q)$ and/or $tr(o_p,o_q)$ values. Then we apply the same idea as $\succsim_\alpha^{\mathrm{L}}$ to obtain a refined dominance relation using $ir(o_p,o_q)$ or $tr(o_p,o_q)$. Namely, we obtain

$$\begin{aligned}\succsim_\tau^{\mathrm{ir}}=\{(o_p,o_q)\mid{}&ir(o_p,o_q)\le\tau \text{ and } (\forall\zeta<\tau,\ \forall(o_r,o_s)\in \mathrm{Trcl}(\succsim_\zeta^{\mathrm{ir}}\cup\{(o_p,o_q)\}),\\ &ir(o_r,o_s)\le\tau \text{ or } (o_s,o_r)\notin\mathrm{Trcl}(\succsim_\zeta^{\mathrm{ir}}))\},\end{aligned} \tag{31}$$

$$\begin{aligned}\succsim_\upsilon^{\mathrm{tr}}=\{(o_p,o_q)\mid{}&tr(o_p,o_q)\le\upsilon \text{ and } (\forall\zeta<\upsilon,\ \forall(o_r,o_s)\in \mathrm{Trcl}(\succsim_\zeta^{\mathrm{tr}}\cup\{(o_p,o_q)\}),\\ &tr(o_r,o_s)\le\upsilon \text{ or } (o_s,o_r)\notin\mathrm{Trcl}(\succsim_\zeta^{\mathrm{tr}}))\}.\end{aligned} \tag{32}$$

Taking their transitive closures, we obtain weak orders among alternatives. τ and υ are defined by the minimum values such that their trensitive clusures become weak orders.

Example 3. Consider the same pairwise comparison matrix and normalized interval weight vector $\boldsymbol{W}$ as in Example 1. We obtain $ir(o_p,o_q)$ and $tr(o_p,o_q)$ as shown in Tables 5 and 6. Then, with $\tau=0.339898$ and $\upsilon=0.463964$, we obtain

$$\begin{aligned}\succsim_\tau^{\mathrm{ir}}=\{&(o_1,o_4),(o_3,o_4),(o_5,o_4),(o_2,o_4),(o_1,o_5),\\ &(o_1,o_3),(o_1,o_2),(o_2,o_5),(o_2,o_3),(o_5,o_3)\},\end{aligned} \tag{33}$$

$$\begin{aligned}\succsim_\upsilon^{\mathrm{tr}}=\{&(o_1,o_4),(o_3,o_4),(o_5,o_4),(o_2,o_4),(o_1,o_2),\\ &(o_1,o_3),(o_2,o_3),(o_5,o_2),(o_5,o_1)\}.\end{aligned} \tag{34}$$

Eventually, we obtain refined weak orders for $\tau\ge 0.339898$ and $\upsilon\ge 0.463964$, $o_1\,\hat{\succsim}_\tau^{\mathrm{ir}}\,o_2\,\hat{\succsim}_\tau^{\mathrm{ir}}\,o_5\,\hat{\succsim}_\tau^{\mathrm{ir}}\,o_3\,\hat{\succsim}_\tau^{\mathrm{ir}}\,o_4$ and $o_5\,\hat{\succsim}_\upsilon^{\mathrm{tr}}\,o_1\,\hat{\succsim}_\upsilon^{\mathrm{tr}}\,o_2\,\hat{\succsim}_\upsilon^{\mathrm{tr}}\,o_3\,\hat{\succsim}_\upsilon^{\mathrm{tr}}\,o_4$, where $\hat{\succsim}_\tau^{\mathrm{ir}}$ and $\hat{\succsim}_\upsilon^{\mathrm{tr}}$ are transitive closures of $\succsim_\tau^{\mathrm{ir}}$ and $\succsim_\upsilon^{\mathrm{tr}}$, respectively.

4 Concluding Remarks

As shown in Example 1, the dominance relation obtained by the conventional AHP is not always unswerving even when the given pairwise comparison matrix is

Table 5: $ir(o_p, o_q)$

	o_1	o_2	o_3	o_4	o_5
o_1	–	0.302326	0.275643	0	0.119997
o_2	0.507693	–	0.332965	0.0890831	0.325053
o_3	0.421277	0.360001	–	0	0.339984
o_4	1	0.5	1	–	1
o_5	0.382839	0.347844	0.339898	0	–

Table 6: $tr(o_p, o_q)$

	o_1	o_2	o_3	o_4	o_5
o_1	–	0.337131	0.409472	0	0.536036
o_2	0.662869	–	0.412291	0.193388	0.538855
o_3	0.590528	0.587709	–	0	0.466514
o_4	1	0.806612	1	–	1
o_5	0.463964	0.461145	0.533486	0	–

sufficiently consistent. We showed that various weak orders are obtained depending on the idea of refinement of dominance relation. In ranking alternatives, those possible weak orders should be considered and the dominance relations obtained by the proposed approach should be interpreted in the real world setting. Moreover, we may combine the proposed tolerance and reduction approaches. These would be included in future topics.

References

[1] Pairwise comparison based interval analysis for group decision aiding with multiple criteria. *Fuzzy Sets and Systems*, 274:79 – 96, 2015.

[2] M. Inuiguchi and S. Innan. Comparison among several parameter-free interval weight estimation methods from a crisp pairwise comparison matrix. *The CD-ROM Proceedings of the 14th International Conference on Modeling Decisions for Artificial Intelligence*, pages 61–76, 2017.

[3] T. L. Saaty. *The Analytic Hierarchy Process*. McGraw-Hill, New York, 1990.

[4] T. L. Saaty and L. G. Vargas. Comparison of eigenvalue, logarithmic least squares and least squares methods in estimating ratios. *Mathematical Modelling*, 5(5):309 – 324, 1984.

[5] K. Sugihara and H. Tanaka. Interval evaluations in the analytic hierarchy process by possibility analysis. *Computational Intelligence*, 17(3):567–579, 2002.

Comparison of Shenoy's Expectation Operator with Probabilistic Transforms and Perez' Barycenter

Radim Jiroušek
Faculty of Management
University of Economics, Prague
radim@utia.cas.cz

Václav Kratochvíl
ÚTIA
Czech Academy of Sciences
velorex@utia.cas.cz

Abstract

Shenoy's paper published in this Proceedings of WUPES 2018 introduces an operator that gives instructions how to compute an expected value in the Dempster-Shafer theory of evidence. Up to now, there was no direct way to get the expected value of a utility function in D-S theory. If needed, one had to find a probability mass function corresponding to the considered belief function, and then - using this probability mass function - to compute the classical probabilistic expectation.

In this paper, we take four different approaches to defining probabilistic representatives of a belief function and compare which one yields to the best approximations of Shenoy's expected values of various utility functions. The achieved results support our conjecture that there does not exist a probabilistic representative of a belief function that would yield the same expectations as the Shenoy's new operator.

1 Introduction

Criteria for finding optimal decisions are usually based on a maximum expected utility principle. As Glenn Shafer [8] wrote already in 1986: *The controversy raised by this book* (here he meant the Savage's book [7]) *and Savage's subsequent writings is now part of the past. Many statisticians now use Savage's idea of personal probability in their practical and theoretical work, ... To do otherwise is to violate a canon of rationality.* This reflects the fact that the maximum expected utility principle is often used not only when the knowledge from the respective field of application is embodied in a probabilistic model but also when the applied model is built within the framework of belief function theory. Nevertheless, to compute the necessary value of expected utility, the respective belief function is usually transformed into an appropriate probability distribution. For this, several procedures

were designed - we call them *probability transforms* in this paper. As advocated by Cobb and Shenoy, the only one, which is compatible with the Dempster-Shafer theory of belief functions is the plausibility transform [1]. The other transforms are more likely compatible with the theory of belief functions interpreted as *generalized probability* [4]. This interpretation reflects the fact that a belief function specifies a convex set of probability distributions, which is called a credal set. In this paper we consider widely used *pignistic transform* advocated by Philippe Smets [10], and two others that are usually omitted in the context of belief function: *maximum entropy* and *Perez' barycenter* [6].

To our best knowledge, the first idea how to compute an expected value for a belief function directly, i.e., avoiding its transformation into a probability distribution, is due to Prakash Shenoy [9]. From the theoretical point of view, it is a concept deserving a deep further investigation. As we will see in the following paragraph, it is defined with the help of commonality functions, which means that it suffers from a great computational complexity. If new computational procedures (avoiding the calculation of a commonality function and subsequent summation over all nonempty subsets of a state space) are not found, the application of this approach in practical problems will be limited. Though we conjecture that there does not exist a probability transform that would yield the same expectations as the Shenoy's operator, there arises an interesting problem: find a probability transform, which approximates the results of the new operator best. And it is the goal of this paper to compare the above-mentioned four probability transforms from this point of view.

To achieve this goal, the rest of the paper is organized as follows. Section 2 recalls basic concepts of belief function theory and introduces the necessary notation. In Section 3, four selected probability transforms are formally introduced. A battery of basic assignments, as well as a set of utility functions used for comparison are presented in Section 4. The main result of this paper (the comparison of the computed expected values) is presented in Section 5. The paper is concluded by Section 6, where the further research is proposed.

2 Notation

Suppose X is a random variable with a finite state space Ω_X. Let 2^{Ω_X} denote the set of all *non-empty*[1] subsets of Ω_X. A *basic probability assignment* (basic assignment for short) m for X is a function $m : 2^{\Omega_X} \to [0, 1]$ such that

$$\sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) = 1.$$

The subsets $\mathsf{a} \in 2^{\Omega_X}$ such that $m(\mathsf{a}) > 0$ are called *focal* elements of m. An important example is the vacuous basic assignment for X, denoted by ι_X, such that $\iota_X(\Omega_X) = 1$. It corresponds to a total ignorance. If all focal elements of m

[1] Notice that we exclude the empty set from 2^{Ω_X} in this paper.

are singletons (one-element subsets) of Ω_X, then we say m is *Bayesian*. In this case, m is equivalent to a probability distribution.

The information in a basic assignment m can be equivalently represented by corresponding *belief* and *plausibility* functions Bel_m and Pl_m, respectively that are defined as

$$Bel_m(\mathsf{a}) = \sum_{\mathsf{b}\in 2^{\Omega_X}\,:\,\mathsf{b}\subseteq\mathsf{a}} m(\mathsf{b}), \qquad Pl_m(\mathbf{a}) = \sum_{\mathbf{b}\in 2^{\Omega}:\,\mathbf{b}\cap\mathbf{a}\neq\emptyset} m(\mathbf{b}),$$

for all $\mathsf{a} \in 2^{\Omega_X}$. In this paper we need also the fourth possibility of expressing a belief function. A *commonality function* for m is defined for all $\mathsf{a} \in 2^{\Omega_X}$

$$Q_m(\mathsf{a}) = \sum_{\mathsf{b}\in 2^{\Omega_X}\,:\,\mathsf{b}\supseteq\mathsf{a}} m(\mathsf{b}).$$

Notice that it is obvious that for all $\mathbf{a} \in 2^{\Omega}$, $Bel(\mathbf{a}) \leq Pl(\mathbf{a})$. For singletons (one-element subsets of Ω_X) commonality and plausibility functions coincide:

$$Q_m(\{x\}) = Pl_m(\{x\})$$

for all $x \in \Omega_X$. Since we consider only normal basic assignments for which $\sum_{\mathsf{a}\in 2^{\Omega_X}} m(\mathsf{a}) = 1$, it can be shown that

$$\sum_{\mathsf{a}\in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} Q_m(\mathsf{a}) = 1.$$

For a basic assignment m on Ω_X and the corresponding commonality function Q_m, Shenoy proposes a new operator computing the expected value of a general function[2] $g : 2^{\Omega_X} \longrightarrow \mathbb{R}$ [9]. Let us adopt his approach to the computation of an expected value of utility function $u : \Omega_X \longrightarrow \mathbb{R}$. First we need to extend the utility function from Ω_X to the whole 2^{Ω_X} (we denote the extension $\hat{u}$) in the way that for all $\mathsf{a} \in 2^{\Omega_X}$

$$\min_{x\in\mathsf{a}}\{u(x)\} \leq \hat{u}(\mathsf{a}) \leq \max_{x\in\mathsf{a}}\{u(x)\}.$$

Following Shenoy's idea we take the weighted average

$$\hat{u}(\mathsf{a}) = \frac{\sum_{x\in\mathsf{a}} u(x) Q_m(\{x\})}{\sum_{x\in\mathsf{a}} Q_m(\{x\})}$$

(in case that $\sum_{x\in\mathsf{a}} Q_m(\{x\}) = 0$ the value $\hat{u}(\mathsf{a})$ does not influence the resulting expected value of u and therefore we can choose any value from the above specified interval; for example $\hat{u}(\mathsf{a}) = (\min_{x\in\mathsf{a}}\{u(x)\} + \max_{x\in\mathsf{a}}\{u(x)\})/2$). Then Shenoy defines the expected value of u with respect to m as follows:

$$E_m(u(X)) = \sum_{\mathsf{a}\in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} \hat{u}(\mathsf{a}) Q_m(\mathsf{a}).$$

[2] $\mathbb{R}$ denotes the set of real numbers.

The last notion introduced in this section was already mentioned in Introduction. Basic assignment m specifies the following convex set of probability distributions P on Ω ($\mathcal{P}_\Omega$ denote the set of all probability distributions on Ω):

$$\mathcal{P}(m) = \left\{ P \in \mathcal{P}_\Omega : \sum_{x \in \mathsf{a}} P(x) \geq Bel_m(\mathsf{a}) \text{ for } \forall \mathsf{a} \in 2^\Omega \right\}.$$

$\mathcal{P}(m)$ is called a *credal set* of basic assignment m. If m is Bayesian, then $\mathcal{P}(m)$ contains just one probability distribution.

3 Probability transforms

In this paper, we study properties of the following four mappings that assign a probability distribution to each basic assignment. For other probability transforms see e.g. [2]. Perhaps, the most famous is *pignistic transform*, defined for all $x \in \Omega_X$ by the formula

$$Bet\_P_m(x) = \sum_{\mathbf{a} \in 2^\Omega : x \in \mathbf{a}} \frac{m(\mathbf{a})}{|\mathbf{a}|}.$$

Another transform is the so-called *plausibility transform*, which is the respective plausibility function normalized on singletons. Formally it is defined for all $x \in \Omega_X$

$$Pl\_P_m(x) = \frac{Pl(\{x\})}{\sum\limits_{y \in \Omega_X} Pl(\{y\})}.$$

The other two probability transforms select a specific representative from the corresponding credal set. One is the *Maximum entropy* element of $\mathcal{P}(m)$, i.e.,

$$Me\_P_m(x) = \arg \max_{P \in \mathcal{P}(m)} H(P),$$

where $H(P)$ is the Shannon entropy of probability distribution P

$$H(P) = -\sum_{x \in \Omega_X} P(x) \log_2 P(x).$$

The other is the Perez' barycenter [6] that has undeservedly fallen into oblivion:

$$Bac\_P_m(x) = \arg \min_{P \in \mathcal{P}(m)} \max_{Q \in \mathcal{P}(m)} Div(Q;P),$$

where $Div(Q;P)$ denote the well-known relative entropy (called also Kullback-Leibler divergence in the literature)

$$Div(Q;P) = \begin{cases} +\infty, & \text{if } \exists\, x \in \Omega_X : P(x) > 0 = Q(x); \\ \sum\limits_{x \in \Omega_X} P(x) \log \left(\frac{P(x)}{Q(x)} \right), & \text{otherwise}^3. \end{cases}$$

[3] We always take $0 \log \left(\frac{0}{0}\right) = 0$.

4 Basic assignments and utility functions

All the examples presented in this paper correspond to a situation when a color ball is drawn from an urn. We consider $\Omega_X = \{r, b, y, g, w\}$, and the random variable X achieves its value in correspondence whether the color of a drawn ball is red, blue, yellow, green, or white.

Though quite uninteresting from the point of view of this paper (we will see it later), we cannot avoid the vacuous basic assignment ι_X representing a total ignorance. In this case, we do not have any other information about the balls in the urn but

- there is at least one ball in the urn ($\emptyset$ is excluded from 2^{Ω_X});
- the urn contains balls of the specified colors only.

We will also consider a situation described by the famous Ellsberg's example [3]. He considers the situation when the urn contains ninety balls, thirty of them are red, the remaining balls are either blue or yellow with unknown proportion. It may even happen that all of the remaining sixty balls are of the same color – blue or yellow. This situation is well described by a basic assignment m_e with two focal elements: $m_e(\{r\}) = \frac{1}{3}$ and $m_e(\{b, y\}) = \frac{2}{3}$.

Like the Ellsberg's example, a one-red-ball example [5] describes a situation in which the behavior of human decision-makers is considered paradoxical. In this example we know the total number of balls in the urn (it equals n) and that one and only one ball is red. The proportion of the remaining colors in the urn is unknown. The situation is depicted by basic assignment $m_{r,n}$ with two focal elements: $m_{r,n}(\{r\}) = \frac{1}{n}$ and $m_{r,n}(\{b, y, g, w\}) = \frac{n-1}{n}$. In the next section we will consider several such basic assignments with different total numbers of balls. Thus, e.g., for $n = 5$ we will consider $m_{r,5}(\{r\}) = \frac{1}{5}$ and $m_{r,5}(\{b, y, g, w\}) = \frac{4}{5}$.

An interesting situation is got when we consider a basic assignment expressing the knowledge that, like in the Ellsberg's example, only balls of three colors (red, blue, and yellow) are in the urn, and we know that at least 20 % of them are red and not more than 50 % are yellow. This knowledge is expressed by the following basic assignment m_q: $m_q(\{r\}) = 0.2$, $m_q(\{r, b\}) = 0.5$, $m_q(\{r, b, y\}) = 0.3$. Notice that in this case the focal elements of m_q are nested ($\{r\} \subseteq \{r, b\} \subseteq \{r, b, y\}$) , and therefore the corresponding belief function is known to be a possibilistic measure.

Another possibilistic measure is the following basic assignment m_p for which: $m_p(\{r\}) = 0.1$, $m_p(\{r, b\}) = 0.2$, $m_p(\{r, b, y\}) = 0.3$, $m_p(\{r, b, y, g\}) = 0.2$, $m_p(\Omega) = 0.2$.

For a survey of all basic assignments considered in the following section see Table 1. In this table, only focal elements are presented. In other words, if a set $\mathsf{a} \in 2^{\Omega}$ does not explicitly appear in the table, it means that its corresponding basic assignment equals 0.

For the purpose of this paper, we used just eight utility function. Naturally, to make a really serious comparison of probability functions we expect to use much larger batteries of basic assignments and utility functions, as well as we expect

Table 1: Basic assignments

denotation	values of all focal elements
ι_X	$\iota_X(\Omega) = 1$
m_e	$m_e(\{r\}) = \frac{1}{3}$, $m_e(\{b, y\}) = \frac{2}{3}$
$m_{r,n}$	$m_{r,n}(\{r\}) = \frac{1}{n}$, $m_{r,n}(\{b, y, g, w\}) = \frac{n-1}{n}$
m_q	$m_q(\{r\}) = 0.2$, $m_q(\{r, b\}) = 0.5$, $m_q(\{r, b, y\}) = 0.3$
m_p	$m_p(\{r\}) = 0.1$, $m_p(\{r, b\}) = 0.2$, $m_p(\{r, b, y\}) = 0.3$, $m_p(\{r, b, y, g\}) = 0.2$, $m_p(\Omega) = 0.2$
m_a	$m_a(\{r, b\}) = 0.2$, $m_a(\{y, g, w\}) = 0.3$, $m_a(\Omega) = 0.5$

to widen also the set of the compared probability transforms. For the considered utility functions see Table 2. Notice that the first four utility functions correspond to the Ellsberg's example.

Table 2: Utility functions

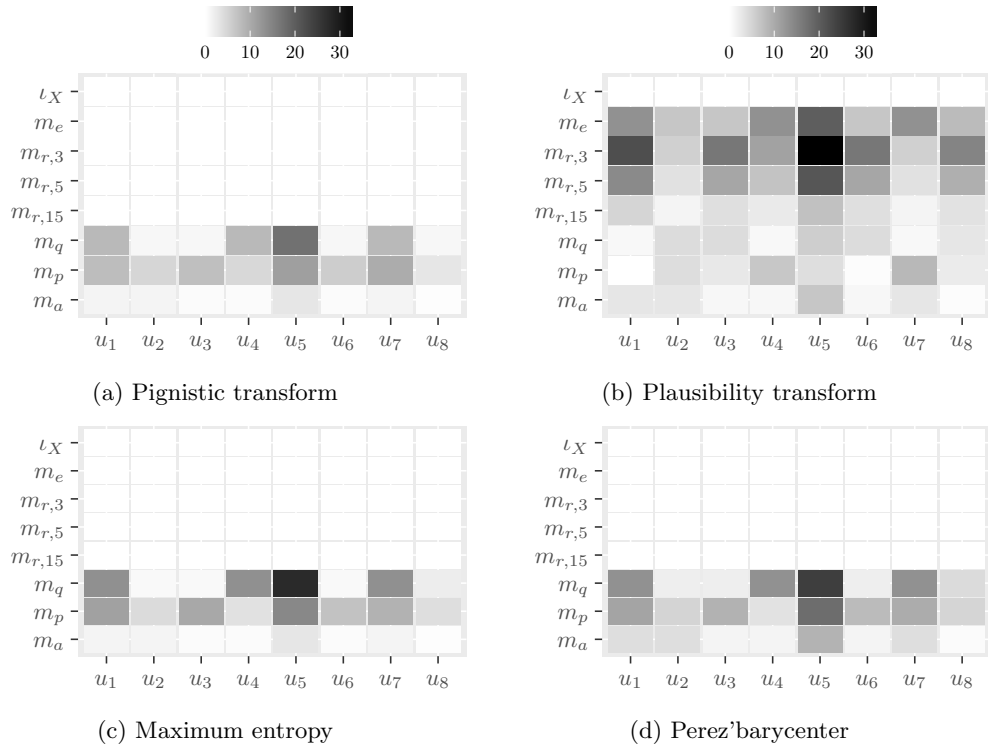
	r	b	y	g	w
u_1	100	0	0	0	0
u_2	0	100	0	0	0
u_3	100	0	100	0	0
u_4	0	100	100	0	0
u_5	0	100	200	300	0
u_6	0	100	0	200	0
u_7	100	0	0	200	100
u_8	50	150	70	220	30

5 Computations

In this section, we describe results obtained from the experimental computations. For each pair, a basic assignment from Table 1 (we considered three basic assignments corresponding to one-red-ball example: $m_{r,3}$, $m_{r,5}$, and $m_{r,15}$, i.e., 8 basic assignments in total) and a utility function from Table 2 we compute five values:

- Shenoy's expected utility value;
- expected utility value computed using pignistic transform;
- expected utility value computed using plausibility transform;

- expected utility value computed using maximum entropy transform;
- expected utility value computed using Perez' barycenter transform.



(a) Pignistic transform

(b) Plausibility transform

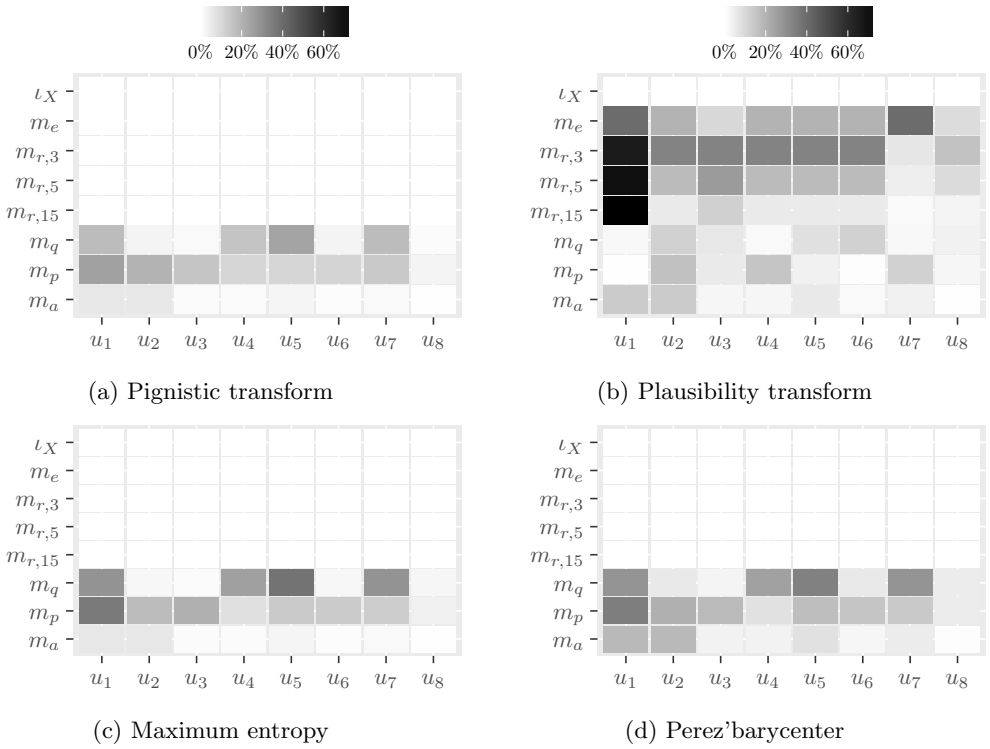
(c) Maximum entropy

(d) Perez'barycenter

Figure 1: Difference between Shenoy's expected utility values and those computed using probability transforms

Each expected utility value computed using a probability transform is then compared with the corresponding Shenoy's expected utility value. Thus, for each probability transform we receive $8 \times 8 = 64$ matrix of values (absolute values of the differences) expressing the difference between the results achieved with the help of the corresponding probability transform and those achieved by the new operator. To make it visually attractive, we depict each such matrix by a 8×8 table, where each difference corresponds to one box. The darker the box, the higher the corresponding difference. Figure 1 depicts the corresponding differences, Figure 2 depicts by how many percent the expected value computed with the help of the respective probability transform differs from the Shenoy's expected value.

We see that the first row in all tables corresponding to ι_X is empty meaning that under the condition of total ignorance all the considered approaches yield



(a) Pignistic transform

(b) Plausibility transform

(c) Maximum entropy

(d) Perez'barycenter

Figure 2: Relative difference between Shenoy's expected utility values and those computed using probability transforms

the same expected utility (all probability transforms give the uniform probability distribution).

6 Conclusions

Though the results achieved in this study should be considered preliminary, they give a hint that the plausibility transform, regardless it is considered by Cobb and Shenoy the only one corresponding to Dempster-Shafer theory of evidence, is quite unsuitable for estimating the expected utility. The question is whether there is any positive result that can be concluded from the described simple study. The achieved results may support the Smets' conviction that the pignistic transform is the best one for decision-making. The results may also suggest that for the situations described by simple basic assignments, the pignistic transform yield the same results as the maximum entropy principle and the Perez' barycenter.

In any case, this study is a starting milestone for further research. From the

theoretical viewpoint, it would be interesting to know whether our conjecture about the nonexistence of a probability transform yielding the same expected values as Shenoy's operator is true or not. From the practical point of view, because of a great computational complexity of the new expectation operator, it is interesting to perform a study similar to the one presented in this paper, but with much greater the number of basic assignments and a higher the number of utility functions.

Acknowledgement

This work was supported by the Czech Science Foundation (project 16-12010S), and by the Czech Academy of Sciences (project MOST-18-04).

References

[1] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–340, 2006.

[2] F. Cuzzolin. On the relative belief transform. *International Journal of Approximate Reasoning*, 53(5):786–804, 2012.

[3] D. Ellsberg. Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75(2):643–669, 1961.

[4] J. Y. Halpern and R. Fagin. Two views of belief: Belief as generalized probability and belief as evidence. *Artificial Intelligence*, 54:3, 1992.

[5] R. Jiroušek and P. P. Shenoy. Ambiguity aversion and a decision-theoretic framework using belief functions. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, pages 1–7. IEEE, 2017.

[6] A. Perez. The barycenter concept of a set of probability measures as a tool in statistical decision. In *4th Int. Vilnius Conf. on Probability Theory and Math. Statistics*, pages 226–228, 1985.

[7] L. J. Savage. *Foundations of Statistics*. Wiley, New York, 1954.

[8] G. Shafer. Savage-revisited. *Statistical Science*, 1(4):463–501, 1986.

[9] P. P. Shenoy. An expectation operator for belief functions in the Dempster-Shafer theory. to appear in Proceedings of WUPES'18.

[10] P. Smets. Decision making in the tbm: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005.

p-Hacking & Exchange of Scientific Information: A Game-Theoretic Approach

Aleš A. Kuběna
Department of Econometrics
Institute of Information Theory and Automation, CAS, Prague
akub@vsup.cz

Abstract

p-Hacking is a bad science practise, when researchers selects statistical hypothesis ex post such that they omits unsignificant results. From the game-theoretic point of view, the exchange of scientific information via publications is a Bayesian game: Each player-experimenter publishes a favourable part of the result of the experiment, but the "denominator" of this result (= number of test and set of negative results) remains his private information. Publishing and citation practice then demotivates researchers to show full and correct results, and favors the p-Hacking-biased results. As a step to the solution, I propose the concept of Compromise Correction. Firstly we adjust the obtained p-values by transforming $p \to \frac{1}{p}$ Now we model the multiple-testing problem as a cooperative game: For each S - subset of the tests, value of characteristic function is $\nu(S) = \underset{i \in S}{\text{Max}} \left(\frac{1}{p_i} \right)$ The idea of compromise correction is to solve the problem of multiple testing, when taking into account the worth of each coalition = each subset of set of experiments. The solution to this problem is the Shapley value:

$$p_k \to \frac{1}{\frac{1}{kp_k} - \sum_{i>k}^{n} \frac{1}{(i(i-1))p_i}}$$

$$p_1 \leq p_2 \leq ... \leq p_n$$

The solution of this problem is tractable, keeps the order of values, is robust w.r.t. changes in "tail" (e.g. increasing the number of strongly negative results) and is piecewise-rational. Robustness w.r.t. tail changes is a property that motivates testing and additional questions and publishes results correctly without fear of relativizing the already achieved significant results. Bonferoni's, resp. Šidak's correction, unlike a compromise correction, limits the maximum "safe" number of additional tests

1 Introduction

1.1 Problem: p-hacking & replication crisis

The replication crisis is a crisis of credibility of the published results of scientific experiments. There is a growing suspicion that many of the results reported as statistically correct, were in actual fact the fluctuation of the random component of observing one experiment in one laboratory. We must always take into account the non-zero frequency of non-replicable and randomly emerged results; part of them arises necessarily, on the basis of statistical error or professional misconduct. However, the ratio of studies whose published results had not been repeated even within the maximum imitation of the original laboratory conditions, significantly exceeds the degree of what could be explained by statistical error. Extensive research [1] has succeeded in repeating 25 % of 67 articles; all of them were from the oncology and haematology areas. This *replication crisis* increases research costs into amounts spent on fruitless follow-up clinical trials and, in addition, threatens health as well as confidence in science. A study of a similar type [10] has also shown "resistance" of non-replicable results to the prestige of the journal : " *The reproducibility of published data did not significantly correlate with journal impact factors, the number of publications on the respective target or the number of independent groups that authored the publications.* " [10]

One of the causes is the so-called p-Hacking [15] The idea of p-Hacking is the hypothesis that non-replicable results arise by the researcher hiding some of the experiment's circumstances. Experimenters conceal multiple-testing and publish, without correction, only those conclusions that have been proven to be significant in the experiment. The p-Hacking hypothesis is statistically testable on large data (= p-values from many articles). [2] statistically proved the non-standard behaviour of the p-value curve around the "magical" threshold p=0.05

However, the replication crisis is not reducible to a mere p-value crisis. The problem would not be solved by replacing a p-value by other statistical indicators in scientific outputs. Cherry-picking can be done based on any statistical indicator. For instance, [3] proved limited replicability of effect size of published results. The p-value is advantageous due to its universality and predictable statistical distribution. For negative results, there is an uniform distribution $U(0,1)$ and, for all the results, a mix of uniform and β-distribution [8]

1.2 Cooperative Game Theory: Basic Definitions

The main of this article is to look to *replication crisis* and *p-hacking* from the game-theoretic point of view. So, this chapter contains basic definitions and concepts of the *cooperative game theory.*

Definition: The pair (Ω, v) is a **cooperative game** (in characteristic function form) if Ω is a finite set of players and $\nu : 2^{\Omega} \to \mathbb{R}$ is a characeristic function that assigns to every coalition $S \subseteq \Omega$ an attainable profit $v(S)$ such that $v(\emptyset) = 0$.

A cooperative game is caled

- **aditive**, if for all $S, T \in 2^\Omega$ with $S \cap T = \emptyset$, $v(S \cup T) = v(S) + v(T)$.
- **monotone**, if for all $T, R \in 2^\Omega$ with $S \subset T$, $\nu(S) \leq \nu(T)$
- **superaditive**, if for all $S, T \in 2^\Omega$ with $S \cap T = \emptyset$, $v(S \cup T) \geq v(S) + v(T)$.
- **subaditive**, if for all $S, T \in 2^\Omega$ with $S \cap T = \emptyset$, $v(S \cup T) \leq v(S) + v(T)$.

Superadditivity implies monotonity, but but monotonity does not imply superadditivity. The game class $maxValueGame[]$ examined in the following section is the set of monotones, but generically subaditive games.

Let $\Gamma = \Gamma(\Omega)$ the set of all cooperative games on Ω and by $\Gamma_1 = \Gamma_1(\Omega)$ the subset of all aditive cooperative games on Ω

Definition: A **value** of games is an operator $\Psi : \Gamma \to \Gamma_1$ s.t. $\Psi \circ \Psi = \Psi$

In particular, we define $\Psi_i(v) := \Psi \circ v(\{i\})$. Clearly, $\Psi \circ v$ is uniquelly determined by the numbers $\Psi_i(v)$.

A special case of the value is the **Shapley value**:

Definition (formula): The **Shapley value** is a value ϕ defined by the formula

$$\phi_i \circ v = \sum_{R \supseteq \{i\}} \frac{\Delta_v(R)}{|R|}$$

where $\Delta_R(v) \in \mathbb{R}$ is a **Harsanyi dividend** of the coalition $R \subseteq \Omega$ defined by

$$\Delta_R(v) = \sum_{T \subseteq R} (-1)^{|R|-|T|} v(T)$$

An alternative, but equivalent definition of the Shaplye value is axiomatic. Shapley theorem [13] proves the existence of a unique game-value operator φ assuming it satisfies the following four axioms:

1. **Linearity**: $\varphi(\alpha v + \beta v') = \alpha\varphi(v) + \beta\varphi(v')$ for all $(\Omega, v), (\Omega, v') \in \Gamma$ and $\alpha, \beta \in \mathbb{R}$
2. **Efficiency**: For all games (Ω, v): $\sum_i \varphi_i(v) = v(\Omega)$
3. **Null-player property**: if $i \in \Omega$ is a *null-player*, i.e. $\forall R \subseteq \Omega\ v(R \cup \{i\}) = v(R)$, then $\varphi_i(v) = 0$
4. **Symmetry** (sometimes called *anonymity*): $\varphi(\rho(i))(\rho \cdot v) = \varphi_i(v)$ for every permutation $\rho \in S_\Omega$ (the function $\rho \cdot v$ is defined by $\rho \cdot v(\rho(R)) := v(R)$) for any $R \subseteq \Omega$)

Axioms 1-4 are independent; in [5] and in [12] are examples of values satisfying any 3 of them and not the 4th.

From the geometric point of view, cooperative game is a point of $\mathbb{R}^{2^\Omega}$ and set of all cooperative games Ω is a $2^{|\Omega|} - 1$ dimensional subspace of the vector space $\mathbb{R}^{2^\Omega}$

From the game-theoretic point of view, cooperative game illustrates an economic situation where a coalition profit or cost depends in general on the involved players in a non-aditive way.

Values of games provide a tool how to evaluate the contibutions of the players. In particular, the Shapley value describes a way how to do it in a *fair* way. Linearity means that *fair* value should be linear. In other words, if the same plaers play two games (v_1, v_2) independently, value of every player should be in sum the same as a value of "join" game $v_1 + v_2$ The second axiom is equivalent to the requirement "The maximum coalition will be formed and its profit will be exactly divided". Null player is a player without any benefit of any coalition; null player property means that value of null player should be 0

The axiom of *symmetry* is an expression of equality of all the participating players. This means that the game-value assigned to them is calculated only from their contributions to the coalitions and does not depend on the particular identity of the player.

Once again from the geometric point of view, a value is an operator $\Psi : \Gamma \to \Gamma_1$. First axiom required that Ψ should be linear, i.e. matrix-representable. Harsanyi dividents $(\Delta_R(v))_{R\subseteq\Omega}$ are coefficients in the *unanimity basis* $(u_S)_{\emptyset\neq S\subseteq\Omega}$

$$u_S(R) = \begin{cases} 1 & S \subseteq R \\ 0 & otherwise \end{cases}$$

$(\Delta_R(v))_{R\subseteq\Omega}$ evaluate a net contribution of coallition R to the total profit $v(\Omega)$

Shapley value of u_S is $\frac{1}{|S|}$ for a members of coalllition S and 0 for non-members. The Shapley value divides the net benefit of each coalition, may be negative, among its members.

2 Compromise correction

2.1 Multiple-testing problem

The research design *one experiment - one atomic result* (one null hypothesis OR one estimeted parameter OR one comparison...) is highly inefficient. So, analysis of experimental data usually tests several hypotheses and estimates several parameters. There are many statistical methods for error-controlling of the experiments with multiple testing: Common known Bonfferoni correction p $\to Np$ and similar Šidak's correction $p \to 1 - (1 - p)^N$ [14], where p is the number of tests. Complex procedures as a [4], [7], [6]. However, there is no incentive mechanism to actually use these procedures, to publish full resut including unsuccessful tests. And simultaneously, each of these procedures rapidly aggravate the score of the basic result when giving more tests. Scientist who only publishes positive results (and conceals negative) is more successful in publishing. And the set of published scientific information is biased. In the sense of [11], the market of scientific informations exchange is poorly designed.

2.2 Compromise correction: idea

The purpose of each correction of p-values is conrolling of probabilities of type I errors (false positives)

The idea of *compromise correction* is to evaluate net contributtion in the sense of cooperative game theory of any test result p_i to the best result $Min_i[p_i]$:

Firstly, we adjust the obtained p-values by transforming them such that the higher value formally means the more convincing results (instead of original ordering *lower value = better results*) $p \to 1/p$. We assume that an individually rational experimenter without interest in credibility, whose primary motivation is to show the outcome as significant as possible, published the most significant result only, without any correction. On the other hand, the rule of multiple testing requires the Bonferoni or another correction. The compromise correction is based on the question of which of the values contributes to the most significant result $Max[\frac{1}{p_i}]$

Let us interpret the problem of the best result as a game over partial results. If the experimenter would only execute a subset S of experiment, his best value vould be $Max_{i\in S}[\frac{1}{p_i}]$. The idea of compromise correction is to solve the problem of multiple testing, when taking into account the worth of each coalition = each subset of set of experiments.The solution to this problem is the *Shapley value*. So we calculate the *Shapley value* for the cooperative $maxValueGame$

$$\text{maxValueGame: } v(S) = \text{Max}\left[X_i = \frac{1}{p_i} : i \in S\right]$$

2.3 Compromise correction: solution

The solution of this problem is tractable: Let $p_1 \leq p_2 \leq ... \leq p_n$, $\boldsymbol{X} = \left(\frac{1}{p_i}\right)_{i=1}^{n}$ Then Shapley value in the coordinate k is

$$\text{Shapley}[\text{maxValueGame}[\mathbf{X}]]_k = \frac{X_k}{k} - \sum_{i>k}^{n} \frac{X_i}{i(i-1)}$$

and compromise correction operator asigns to the k-th best value a corrected value

$$p_k \to \frac{1}{\frac{1}{kp_k} - \sum_{i>k}^{n} \frac{1}{i(i-1))p_i}}$$

Proof: Let $X_1 \geq X_2 \geq ... \geq X_n \geq X_{n+1}$

$$maxValueGame[X_1...X_{n+1}] =$$

$$= maxValueGame[X_1 - X_{n+1}, X_2 - X_{n+1}, ...X_n - X_{n+1}, 0] + constantGame[X_{n+1}]$$

where $constantGame[y][S] = y$ for any nonempty coallition S

By linearity, Shapley value of the $maxValueGame[X_1...X_{n+1}]$ = is the sum of Shapley values of two games defined above. For the first game, value 0 is the

nullplayer in the sense of Shapley 3rth axiom: $X_i - X_{n+1} \geq 0$ and $Max[S] = Max[S \cup \{0\}]$ for $S \subseteq \{X_1 - X_{n+1}, X_2 - X_{n+1}, ...X_n - X_{n+1}\}$. For the second game, Shapley value is $(\frac{X_{n+1}}{n+1})_{i=1...n}$ according to the symmetry of Shapley value. So

$$\text{Shapley[maxValueGame}[[(X)_1^n \cup \{X_{n+1}\}]]_k = \frac{X_k}{k} - \sum_{i>k}^{n+1} \frac{X_i}{i(i-1)}$$

and after inverse transform

$$p_k \to \frac{1}{\frac{1}{kp_k} - \sum_{i>k}^{n+1} \frac{1}{i(i-1)p_i}}$$

$\diamondsuit$

3 Properties of compromise correction

3.1 Mathematical & computational

- Compromise correction **keeps the order of values**. The k-th best value remains the k-th best value after correction
- **Piecewise linearity** of the operator $(X_i)_{i=1}^n \to \text{Shapley[maxValueGame[X]]}_i$

  We reassess the data with a more sensitive test, and we assume that only the first result will improve,

  $$p'_1 < p_1 \leq p_2 = p'_2 \leq p_3 ... \leq p_n = p'_n$$

  Then the compromise correction of all the improvements also gives the best result,

  $$\hat{p}'_1 < \hat{p}_1 \leq \hat{p}_2 = \hat{p}'_2 \leq \hat{p}_3 ... \leq \hat{p}_n = \hat{p}'_n$$

  where

  $$\hat{(p_i)}_{i=1}^n = \text{CompromiseCorrection}[((p_i)_{i=1}^n]$$

3.2 Game-theoretic, Reverse-game-theoretic & motivational

- **Robustness against tail change**s: Let's assume that the experimenter has achieved a significant result, but there is still material left to test additional questions with a low likelihood of becoming significant. When honestly applying Bonferoni's correction, it is preferable not to carry out further analyzes. The reason is the risk of destroying existing and confirmed results

- the Bonferoni correction coefficient increase after each new test. Compromise corrections cause only small bounded fluctuations of the first results. Intuitively, we consider the test and publishing of additional results to be collectively rational. Compromise correction is not in contradiction with individual rationality.

- Copromise correction is a **centralized mechanism**. It needs to collect all the data for the calculation. However, for the above reasons, it motivates the publication of a whole set of p-values better than Bonferoni's correction.

- Bonferoni's correction defines the upper limit of Compromise correction.

4 Limitations and future work

Unfortunately, compromise correction is a decentralized mechanism. From the reverse-game-theoretic point of view, compromise correction is a half-solution only. Compromise correction eliminates the fears and some disadvantages of complying with the rules of good science, but it does not remove the temptation to do so. Compromise correction works as a lifeline for researchers willing to publish correctly, but does not work well as a sticks on those who do not care about the replicability of their published results. However, I hope that in the second plane higher number of published full-set results will help to improve of power of detection of inaccuracies. Construction of decentralized mechanism is the next plan of the research.

The second weak point of compromise correction is its implausible behavior in situation with two or more very similar tests, i.e. with tests with a high a-priori conditional probability $P(T_2 \text{ is significant}|T_1 \text{ is significant})$. For instance Kaplan-Meier test and Cox-regression for the same data. The compromise-correction coefficient for both is implausible close to maximal (but still smaller than Bonferroni!). The usual motivation in this case is not to increase the number of tests, but to find out more of the model parameters.

Further research will focus on the ability to replace the Shapley value with value associated with a pre-defined network structure: Myerson value [9]

5 Acknowledgement

This work was supported by the Excelence project P402/12/G097 DYME Dynamic Models in Economics of GAČR

References

[1] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531, 2012.

[2] S. B. Bruns and J. P. Ioannidis. P-curve and p-hacking in observational research. *PLoS One*, 11(2):e0149144, 2016.

[3] O. S. Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[4] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, pages 111–139, 2002.

[5] R. P. Gilles. *The cooperative game theory of networks and hierarchies*, volume 44. Springer Science & Business Media, 2010.

[6] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[7] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[8] L. R. Jager and J. T. Leek. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):1–12, 2013.

[9] R. B. Myerson. On the value of game theory in social science. *Rationality and Society*, 4(1):62–73, 1992.

[10] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712, 2011.

[11] A. E. Roth. *Who Gets What–and why*. Brilliance Audio, 2015.

[12] D. Schmeidler. The nucleolus of a characteristic function game. *SIAM Journal on applied mathematics*, 17(6):1163–1170, 1969.

[13] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[14] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[15] U. Simonsohn, L. D. Nelson, and J. P. Simmons. P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534, 2014.

Comparison and Connection between the joint and the conditional Generalized Iterative Scaling Algorithm

Carlotta Langer
Martin-Luther-University
Halle-Wittenberg
carlotta.langer@student.uni-halle.de

Nihat Ay
MPI for Mathematics in the Sciences
Santa Fe Institute
University of Leipzig
nay@mis.mpg.de

Abstract

Iterative Scaling is a widely used method to solve maximum entropy problems. Depending on the application they are used for, there are many different versions of Iterative Scaling algorithms. This paper compares and reconnects two popular algorithms, which share a name, but are not equal and even converge to different limit points.

Keywords: iterative scaling, maximum entropy estimation

1 Introduction

Beginning with the Iterative Scaling algorithm described by Csiszár in [6], there now exist many different types of Iterative Scaling algorithms depending on the respective application. There are two frequently used algorithms, which converge to different limit points but are both called Generalized Iterative Scaling (GIS) algorithm. These algorithms are the one presented by Darroch and Ratcliff in [8] and the algorithm used for example by Goodman in [10] or Huang et al. in [11]. To prevent confusion we will call the last one Conditional Generalized Iterative Scaling (CGIS) algorithm.

In order to emphasize their similarities and differences we present both algorithms and categorize them according to two different properties. At first we take a look at the instance they are operating on. In the case of the GIS algorithm it is the full probability distribution, in contrast to the CGIS algorithm that operates on parameters λ. We introduce an intermediate algorithm, the *Joint GIS* algorithm to analyse this difference.

Secondly, the form of the probability distribution is different in both algorithms. The GIS and *Joint GIS* algorithm use joint distributions while the CGIS algorithm computes conditional distributions. This transformation is the reason for the different limit points. GIS and *Joint GIS* converge to the maximum entropy estimation, in contrast to CGIS which converges to the conditional maximum entropy estimation. Our tool to study this difference is the *conditional GIS* algorithm, a new algorithm which works on the whole probability distribution but with the conditional maximum entropy principle. The categorization of these four algorithms is shown in Figure 1.

iteration on \ distribution	joint	conditional
full distribution	GIS [8], [5] Section 3	*conditional GIS* Section 5
parameter	*Joint GIS* Section 4	CGIS [10], [11] Section 6

Figure 1: Categorization of the four different algorithms, namely GIS, *conditional GIS*, *Joint GIS* and CGIS according to the used distribution and the component over which they iterate.

2 Iterative Scaling

This section gives a brief introduction to Iterative Scaling. The probability distributions discussed here are discrete distributions on a finite set X and the set of these distributions will be denoted by $\mathcal{P}$.

Iterative Scaling algorithms are a method to solve maximum entropy problems. In this setting, the algorithms determine a probability distribution on X with predetermined properties. The properties are fixed by introducing constraints which describe the expected value of a feature and are defined for $i \in \{1, \dots, m\}$ as:

$$\sum_{x \in X} P(x) f_i(x) = k_i, \quad k_i \geqslant 0, \ \sum_i k_i = 1. \tag{1}$$

The constraints are called consistent, if the set of positive probability distributions on X, which fulfil these, is not empty. An important result for Iterative Scaling is the duality in Lemma 1. We will need the following sets:

$$L(f,k) = \{P \in \mathcal{P} \mid \sum_{x \in X} P(x) f_i(x) = k_i, \ i \in \{1, \dots, m\}\}$$

$$Q(f, P^{(0)}) = \left\{ P \in \mathcal{P} \mid P(x) = \frac{1}{Z_{P^{(0)}}(\lambda \cdot f)} e^{\sum_i \lambda_i f_i(x)} P^{(0)}(x), \ \lambda_i \in \mathbb{R}, \ x \in X \right\}.$$

Lemma 1. Suppose that the distribution $\widehat{P}$ satisfies the constraints and that $D(\widehat{P} \parallel P^{(0)}) < \infty$ for a probability distribution $P^{(0)}$. Then any of the following properties determine $P^\star$ uniquely and the following statements are equivalent:

(1) $P^\star = \underset{\widehat{Q} \in \overline{Q}(f, P^{(0)})}{\arg\min} \; D(\widehat{P} \parallel \widehat{Q})$

(2) $P^\star = \underset{P \in L(f,k)}{\arg\min} D(P \parallel P^{(0)})$

(3) $P^\star \in L(f,k) \cap \overline{Q}(f, P^{(0)})$

Proof. The proof is given by Ay et al. in Section 2.8.3 Theorem 2.8 in [1]. □

3 The GIS algorithm

In order to prove the convergence of the algorithm Darroch and Ratcliff apply in [8] the following restrictions to the features:

$$\sum_x f_i(x) = 1 \quad \text{and} \quad f_i(x) \geqslant 0. \tag{2}$$

It is possible to define less strict restrictions as for example Curran and Clark point out in [7], but at this point it is sufficient to use the original ones.

Theorem 3.1 (The GIS algorithm, [8])**.** Let $P^{(0)}$ be the uniform distribution, f_i as in (2), $n \in \mathbb{N}$ and

$$P^{(n)}(x) = P^{(n-1)}(x) \prod_{i=1}^{m} \left(\frac{k_i}{\sum\limits_{x' \in X} P^{(n-1)}(x') f_i(x')} \right)^{f_i(x)}.$$

The $P^{(n)}$ converges to a positive and unique solution $P^\star \in \overline{Q}(f, P^{(0)})$ fulfilling the constraints (1) and the properties described in Lemma 1.

Additionally, it is possible to prove that $P^\star \in L(f,k) \cap Q(f, P^{(0)})$ under the condition that the constraints are consistent.

Proof. This was proven by Csiszár in [5]. □

4 The *Joint GIS* algorithm

The parameters λ_i and the features f_i determine the density of a Gibbs distribution uniquely. Therefore, we do not have to compute $P^{(n)}$ in each step of the algorithm in order to find $P^\star$. That is why it is sufficient to find an iteration for the parameters. In each step, the parameters λ_i will be altered by δ_i, so that $\lambda_i^{(n+1)} = \lambda_i^{(n)} + \delta_i$.

Although it is easy to check that the iteration below and the one described in Theorem 3.1 are related, we are not able to conclude immediately that this new algorithm converges. Darroch and Ratcliff state in [8] at the end of Section 2 without a proof that the parameters can be compiled in an easy way. In [2] Brown et al. are able to give this proof for some cases, but not for the general case.

In conclusion, we will prove the convergence of the algorithm based on the framework Pietra et al. use in [13] to prove the convergence of another form of iterative scaling algorithm, their Improved Iterative Scaling algorithm. Denote by

$$P_\delta(x) = \frac{1}{Z(\lambda+\delta, f)} \cdot e^{\sum_{i=1}^{m} (\lambda_i+\delta_i) f_i(x)}, \; Z(\lambda+\delta, f) = \sum_{x \in X} e^{\sum_{i=1}^{m} (\lambda_i+\delta_i) f_i(x)}. \tag{3}$$

The goal is to maximize $M(Q, P_\delta) := -D(Q \parallel P_\delta)$ with a fixed $Q \in \mathcal{P}$ that satisfies the constraints and to use Lemma 1. Therefore we need to find a lower bound of the steps of the iteration which is easy to maximize respecting δ. The following Definition 1 and Theorem 4.1 are the ones Pietra et al. use in [13] in Section 4.B.

Definition 1. A function $B : \mathbb{R}^m \times \mathcal{P} \to \mathbb{R}$ is an auxiliary function for $M(Q, P_\delta)$ if it holds the following properties:

(1) For all $P \in \mathcal{P}$ and $\delta \in \mathbb{R}^m$ we have: $M(Q, P_\delta) \geqslant M(Q, P) + B(\delta, P)$.

(2) $B(\delta, P)$ is continuous in $P \in \mathcal{P}$ and C^1 in $\delta \in \mathbb{R}^m$.

(3) Let $t \in \mathbb{R}$. Then $B(0, P) = 0$ and:

$$\left.\frac{d}{dt}\right|_{t=0} B(t \cdot \delta, P) = \left.\frac{d}{dt}\right|_{t=0} M(Q, P_{t\cdot\delta}).$$

It is possible to define the following sequence:

$$P^{(n+1)} = P^{(n)}_{\delta^{(n)}}, \text{ with } \delta^{(n)} = \arg\max_{\delta \in \mathbb{R}^m} B(\delta, P^{(n)}).$$

Property (1) of Definition 1 makes sure that $M(Q, P_{\delta^{(n)}})$ increases with every step. With this we can get to the next result:

Theorem 4.1. Let $P^{(n)} \in \mathcal{P}$ be a sequence where the support of $P^{(0)}$ is X and the properties

$$P^{(n+1)} = P^{(n)}_{\delta^{(n)}}, \; \delta^{(n)} \in \mathbb{R}^{(m)}, \quad B(\delta^{(n)}, P^{(n)}) = \sup_{\delta \in \mathbb{R}^m} B(\delta, P^{(n)}).$$

Then $M(Q, P_{\delta^{(n)}})$ increases monotonically, it converges to

$$\max_{\widehat{Q} \in \overline{Q}(f, P^{(0)})} M(Q, \widehat{Q}) \text{ and } \lim_{n\to\infty} P^{(n)} = P^\star = \arg\max_{\widehat{Q} \in \overline{Q}(f, P^{(0)})} M(Q, \widehat{Q}).$$

Proof. This is proven in [13] Section 4.B. □

In order to resemble the CGIS algorithm, we will now make use of new restrictions towards the features:

$$f^c := \max_{x\in X} \sum_{i=1}^{m} f_i(x) \quad 0 \leqslant f_i(x) \leqslant 1, \quad i = 1, \ldots, m.$$

Additionally, we assume that $f^c \geqslant 1$. Note that we no longer need the f_i to sum up to 1 or any fixed constant. In [7] Curran and Clark proved the convergence of the algorithm without a correction feature

$$f_{m+1} := 1 - \sum_{i=1}^{m} \frac{f_i(x)}{f^c}$$

by fixing its value with $\lambda_{m+1} \equiv 0$ to zero. In order to apply the Jensen's inequality we will use the same trick in the next lemma.

Lemma 2. Let $Q, P \in \mathcal{P}$ and $\delta \in \mathbb{R}^m$. Then the function

$$B(\delta, P) = 1 + \sum_{x\in X} Q(x) \sum_{i=1}^{m} \delta_i f_i(x) - \sum_{x\in X} P(x) \sum_{i=1}^{m+1} \frac{f_i(x)}{f^c} e^{\delta_i f^c}$$

is an auxiliary function for $M(Q, P)$ with $\delta_{m+1} = 0$ fixed.

Proof of Lemma 2. We will prove the properties listed in Definition 1. To prove the first property we use $log(x) \leqslant x - 1$, for all $x > 0$ and the Jensen's inequality.

$$\begin{aligned}(1)\ M(Q, P_\delta) - M(Q, P) &\geqslant \sum_{x\in X} Q(x) \sum_{i=1}^{m} \delta_i f_i(x) - \sum_{x\in X} Q(x) \left(\sum_{x'\in X} e^{\sum_{i=1}^{m} \delta_i f_i(x')} P(x') - 1 \right) \\ &\geqslant B(\delta, P).\end{aligned}$$

(2) The definition of f^c assures that $f^c > 0$. As a sum of continuous functions $B(\delta, P)$ is continuous in P and

$$\frac{d}{d\delta} B(\delta, P) = \left(\sum_{x\in X} Q(x) f_1(x) - P(x) f_1(x) e^{\delta_1 f^c} \ \ldots \ \sum_{x\in X} Q(x) f_m(x) - P(x) f_m(x) e^{\delta_m f^c} \right).$$

Every entry of the Jacobian matrix is continuous in δ_i and we gain property (2).

(3) For $\delta = 0^{(m)}$ we get: $B(0, P) = 1 - \sum_{x\in X} P(x) \sum_{i=1}^{m+1} \frac{f_i(x)}{f^c} = 1 - 1 = 0$. With $t \in \mathbb{R}$ the differentiation leads to:

$$\begin{aligned}\left.\frac{d}{dt}\right|_{t=0} M(Q, P_{t\cdot\delta}) &= \left.\frac{d}{dt}\right|_{t=0} \sum_{x\in X} ln \left(\frac{1}{Z_P(t\cdot\sigma, f)} e^{\sum_{i=1}^{m} t\cdot\delta_i f_i(x)} P(x) \right) - \sum_{x\in X} Q(x) ln(Q(x)) \\ &= \sum_{x\in X} Q(x) \sum_{i=1}^{m} \delta_i f_i(x) - \sum_{x\in X} P(x) \sum_{i=1}^{m} \delta_i f_i(x) = \left.\frac{d}{dt}\right|_{t=0} B(t\cdot\delta, P).\end{aligned}$$

□

It remains to show that $B(Q, P)$ and Theorem 4.1 result in the desired iteration:

Lemma 3 (The *Joint GIS* algorithm)**.** Let $\lambda_i^{(0)} = 0$ and

$$\lambda_i^{(n+1)} = \lambda_i^{(n)} + \frac{1}{f^c} ln \left(\frac{k_i}{\sum\limits_{x \in X} P^{(n)}(x) f_i(x)} \right). \tag{4}$$

This converges to $\lambda_i^\star = \lim\limits_{n \to \infty} \lambda_i^{(n)}$. Additionally we have

$$P(x) = \frac{1}{Z(\lambda^\star, f)} e^{\sum\limits_{i=1}^{m} \lambda_i^\star f_i(x)} = P^\star(x)$$

for all x and $P^\star$ is the same limit point as the one of GIS in Theorem 3.1.

Proof of Lemma 3. Theorem 4.1 provides us with an iteration that convergences to $\arg\max\limits_{\widehat{Q} \in \overline{Q}(f, P^{(0)})} M(Q, \widehat{Q})$. Choosing $\lambda_i^{(0)}$ to be zero leads to $P^{(0)}$ as the uniform distribution. That means that the initial points of both algorithms are the same. Now we will take a look at $\delta^{(n)}$ defined in Theorem 4.1 and maximize $B(\delta, P^{(n)})$ in respect to δ. For every $i \in \{1, \ldots, m\}$ we get:

$$\delta_i = \frac{1}{f^c} ln \left(\frac{k_i}{\sum\limits_{x \in X} P^{(n)}(x) f_i(x)} \right).$$

This is in fact a maximum of $B(\delta, P^{(n)})$ because of the negativity of the Hessian matrix. We are now able to iterate over the λ_i separately, because of the following equality:

$$\sup_{\delta \in \mathbb{R}^m} B(\delta, P^{(n)}) = 1 + \sum_{i=1}^{m} \sup_{\delta_i \in \mathbb{R}} \left(\sum_{x \in X} Q(x) \delta_i f_i(x) - \sum_{x \in X} P(x) \frac{f_i(x)}{f^c} e^{\delta_i f^c} \right).$$

Together with (3) this leads to the iteration stated in this Lemma. This proves the convergence of the algorithm. Lemma 1 additionally yields the equality of the limit of this algorithm and the one of the algorithm described in Theorem 3.1. □

5 The *conditional GIS* algorithm

In this section we will perform the second step towards the implemented CGIS algorithm. This approach was also described in [14] in Section 4.5 and in [3] in Section 3. We will switch from joint distributions to conditional distributions. The GIS algorithm can be very expensive regarding the needed time for each step in the iteration. In each step the algorithm iterates over every $x \in X$. Consider an

experiment as an application for the algorithm with a huge space X of possible outcomes. It is likely to assume that there are applications in which the space of actually occurring x is a rather small subset $X' \subset X$. By substituting the joint probabilities with a special form of probability, we are able to iterate only over the x that actually appear in the data. First, we will introduce this new form of probability distributions regarding a target distribution P_t satisfying the constraints in general. This approach allows us to introduce the next step without assuming the existence of data.

In contrast to the parameter estimation this step actually changes the limit of the convergence. By using a different form of probability distributions the maximum entropy principle turns into the conditional maximum entropy principle.

To do so, we have to be able to write X as $X = X_1 \times X_2$ by defining two disjoint subsets A_1, A_2 of $\{1, \dots, r\}$ with $A_1 \cup A_2 = \{1, \dots, r\}$ and the alphabets $\mathcal{A}_\beta,\ \beta \in \{1, \dots, r\}$ of $x_\beta \in \mathcal{A}_\beta$:

$$X_i = \{x_i = (x_\beta)_{\beta \in A_i} \mid x \in \bigtimes_{\beta \in A_i} \mathcal{A}_\beta\},\ i \in \{1, 2\}.$$

The probability of $P \in \mathcal{P}$ on X_1 is defined by:

$$P(x_{A_1}) = \sum_{x \in X(x_{A_i})} P(x),\ x_{A_1} \in X_1 \text{ with } X(x_{A_i}) = \{y \in \bigtimes_{\beta=1}^{r} \mathcal{A}_\beta \mid y_{A_i} = x_{A_i}\}.$$

With the additional restriction that the marginal possibility of the $x_1 \in X_1$ equals the empirical distribution $\widehat{P}(x_1),\ x_1 \in X_1$ derived from a fixed set of data, we are able to define an algorithm iterating only over the $x_1 \in X_1$ occurring in the tests. Suppose that $P_t \in \mathcal{P}$ satisfies the constraints (1). Now we are able to change $P(x_1)$ to $P_t(x_1)$ to create the new constraints:

$$\sum_{x_1 \in X_1} P_t(x_1) \sum_{x_2 \in X_2} P(x_2 \mid x_1) f_i(x_1, x_2) = k_i. \tag{5}$$

This leads to the definition of a new probability distribution on X:

$$P^c(x) = P_t(x_1) \cdot P(x_2 \mid x_1), \quad \text{for all } x_1 \in X_1,\ x_2 \in X_2.$$

Now we take a closer look at the new probability distribution P^c. While P_t is fixed to the target distribution, $P(x_2 \mid x_1)$ is a conditional Gibbs-distribution:

$$P(x_2 \mid x_1) = \frac{1}{Z_{x_2}(x_1)} e^{\sum_{i=1}^{m} \lambda_i f_i(x_1, x_2)},\ Z_{x_2}(x_1) = \sum_{x_2 \in X_2} e^{\sum_{i=1}^{m} \lambda_i f_i(x_1, x_2)}. \tag{6}$$

It is possible to define similar sets to the ones in Section 2:

$$L^c(f, k) = \left\{ P \in \mathcal{P} \mid P(x_1) = P_t(x_1), \text{ for all } x_1 \in X_1 \text{ and } \sum_{x \in X} P(x) f_i(x) = k_i \right\}$$

$$Q^c(f, P^{(0)}) = \left\{ P \in \mathcal{P} \mid P(x) = P_t(x_1) \frac{1}{Z_{x_2}(x_1)} e^{\sum_{i=1}^{m} \lambda_i f_i(x)} P^{(0)}(x),\ \lambda_i \in \mathbb{R},\ x \in X \right\}.$$

Now we are able to define a conditional equivalent to Lemma 1:

Lemma 4. Suppose that the distribution $\widehat{P} \in L^c(f,k)$ satisfies the constraints and that $D(\widehat{P},U) < \infty$. If $P^\star \in L^c(f,k) \cap Q^c(f,U)$ exists, it is unique and holds the properties:

(1) $P^\star = \underset{\widehat{Q} \in Q^c(f,U)}{\arg\min}\ D(\widehat{P}(X_2 \mid X_1) \parallel \widehat{Q}(X_2 \mid X_1))$

(2) $P^\star = \underset{P \in L(f,k)}{\arg\min}\ D(P(X_2 \mid X_1) \parallel U(X_2 \mid X_1))$

Proof. With the chain rule (9) the proof can be easily derived from the one presented by Pietra et Al. in Proposition 4 in [13]. □

All things considered, we are able to maximize the conditional entropy by adjusting the parameters λ_i of (6). Now we define the *conditional GIS* and it converges to the distribution of the form of P^c that maximizes the conditional entropy.

Lemma 5 (The conditional GIS algorithm)**.** Let k_i be defined as in (5) and $P^{(0)}$ be the uniform distribution. If a probability distribution of the form (6) satisfying the constraints exists, then

$$P^{(n)}(x) = P^{(n-1)}(x) \prod_{i=1}^{m} \left(\frac{k_i}{\sum\limits_{x_1 \in X_1} P_t(x_1) \sum\limits_{x_2 \in X_2} P^{(n-1)}(x_2 \mid x_1) f_i(x_1, x_2)} \right)^{\frac{f_i(x)}{f^c}}$$

converges to $P^\star$ of the form (6), satisfying the constraints and maximizing the conditional entropy.

Proof. The following proof is derived from the one Darroch and Ratcliff presented in [8] of Theorem 1. At first we will use the inequality between the generalized arithmetic and geometric means:

$$\prod_{i=1}^{m} \left(\frac{k_i}{k_i^{(n-1)}} \right)^{\frac{f_i(x)}{f^c}} \leqslant \sum_{i=1}^{m} \frac{f_i(x)}{f^c} \cdot \left(\frac{k_i}{k_i^{(n-1)}} \right) \tag{7}$$

with $k_i^{(n-1)} = \sum\limits_{x_1 \in X_1} P_t(x_1) \sum\limits_{x_2 \in X_2} P^{(n-1)}(x_2 \mid x_1) f_i(x_1, x_2)$. Applying this leads to

$$\sum_{x \in X} P^{(n)}(x) \leqslant \sum_{x \in X} P^{(n-1)}(x) \sum_{i=1}^{m} \frac{f_i(x)}{f^c} \cdot \left(\frac{k_i}{\sum\limits_{x \in X} f_i(x) P^{(n-1)}(x)} \right) = \frac{1}{f^c} \sum_{i=1}^{m} k_i = \frac{1}{f^c}$$

and we have $\sum_{x\in X} P^{(n)}(x) \leqslant \frac{1}{f^c} \leqslant 1$, therefore $\sum_{i=1}^{m}\sum_{x\in X} P^{(n)}(x)f_i(x) \leqslant \frac{1}{f^c} \leqslant 1$. For all n, $P^{(n)}(x) > 0$ and $k_i^{(n)} = \sum_{x\in X} P^{(n)}(x)f_i(x) > 0$. The positivity of the KL-divergence leads to

$$D(k_i \parallel k_i^{(n)}) = \sum_{i=1}^{m} k_i log_2\left(\frac{k_i}{k_i^{(n)}}\right) \geqslant 0. \tag{8}$$

Let Q be an arbitrary probability distribution satisfying (1). Then

$$D(Q \parallel P^{(n+1)}) = \sum_{x\in X} Q(x) log_2\left(\frac{Q(x)}{P^{(n+1)}(x)}\right) = D(Q \parallel P^{(n)}) - \frac{1}{f^c} D(k \parallel k^{(n)}).$$

Now $\{D(Q \parallel P^{(n+1)}), n \in \mathbb{N}\}$ is a decreasing bounded sequence. Therefore it has a limit point and $D(k \parallel k^{(n)}) \to 0$ as $n \to \infty$. The properties of k_i and Pinsker's inequality

$$D(k \parallel k^{(n)}) \geqslant \frac{1}{2}\sum_{i=1}^{m} \mid k_i - k_i^{(n)} \mid^2$$

lead to $k_i^{(n)} \to k_i$ as $n \to \infty$. Suppose P_1, P_2 are different limit points of the bounded sequence $\{P^{(n)}\}$. Because of (5), both satisfy the constraints (1) and are of the form (6). Additionally P_1, P_2 are positive probability distributions. Now we are able to apply Lemma 4 and this yields that $P_1 = P_2$. □

6 The CGIS algorithm

Applying both changes, the parameter estimation and the conditional distribution, leads to the desired CGIS algorithm:

Lemma 6 (The CGIS algorithm, [10], [11], [7])**.** Let $\lambda_i^{(0)} = 0$, k_i be consistent and as in (5) and $P^{(n)}$ of the form (6) with parameters $\lambda^{(n)}$. The iteration

$$\lambda_i^{(n)} = \lambda_i^{(n-1)} + \frac{1}{f^c} ln\left(\frac{k_i}{\sum_{x_1\in X_1} P_t(x_1) \sum_{x_2\in X_2} P^{(n-1)}(x_2 \mid x_1) f_i(x)}\right)$$

converges to the limit $\lambda_i^\star$ with $P^\star(x_2 \mid x_1) = \frac{1}{Z_{x_2}} e^{\sum_{i=1}^{m} \lambda_i^\star f_i(x)}$. Additionally, $P^\star(x) = P_t(x_1)P^\star(x_2 \mid x_1)$ is conditional maximum entropy estimation.

Proof. Curran and Clark provide a proof in the Appendix of [7]. □

7 Comparison

In order to understand the relationship between GIS and CGIS, we will take a look at the following chain rule for entropy:

$$H_P(X_1, X_2) = H_P(X_1) + H_P(X_2 \mid X_1) \tag{9}$$

with the conditional entropy defined as

$$H_P(X_2 \mid X_1) = -\sum_{x_1 \in X_1} P(x_1) \sum_{x_2 \in X_2} P(x_2 \mid x_1) log_2(P(x_2 \mid x_1)).$$

A proof for this rule is given by [4] in Theorem 2.2.1. Notice that the entropy equals the conditional entropy in the case of a fixed marginal distribution on X_1. That means that the algorithms GIS and *Joint GIS* iterate towards the same limit as the conditional ones under the restriction that the marginal distribution on X_1 is fixed in both cases to the same values. This can be easily done by introducing an additional feature for the desired distribution. However, in the general case the limits are not equal, as we can observe in Figure 2 (a). This example for the different limit points of maximum entropy and conditional maximum entropy estimation was given by Yuret in [15]. We gain the values calculated by Yuret with an implementation of *Joint GIS* and CGIS in C++ available at [9].

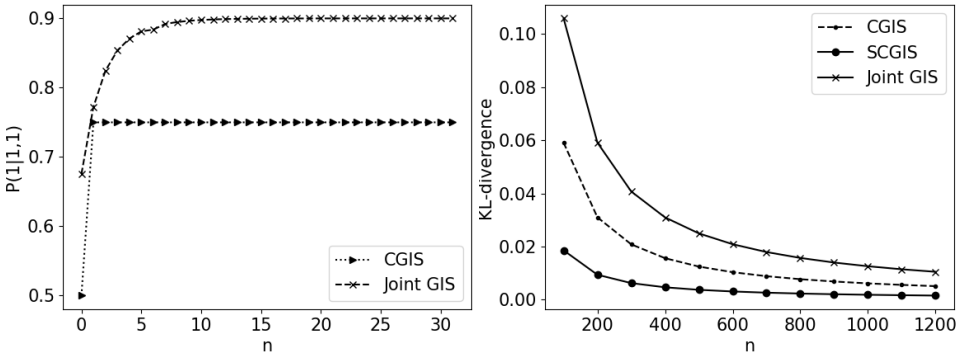


Figure 2: (a) CGIS vs. *Joint GIS* (b) AND: *Joint GIS*, CGIS and SCGIS

Additionally, we are able to compare the performances of the algorithms with an easy example. Consider X as $X = X_1 \times X_2 \times X_3$, with (X_1, X_2) as input and X_3 as output. Now, we are trying to predict the value of X_3 while only knowing the input. Our set of data is the logical AND gate listed in Figure 3(a).

At first we assume that X_3 depends on X_1 and X_2, but not their interactions as illustrated in Figure 3 (a) 2. Computing the data under this assumption leads to an probability distribution, which we compare to a second probability distribution by calculating the KL-divergence between them. The second probability distribution is gained by expecting that X_3 depends on X_1 and X_2 simultaneously and their interaction as visualized in Figure 3 (b) 2.

X_1	X_2	X_3
0	0	0
0	1	0
1	0	0
1	1	1

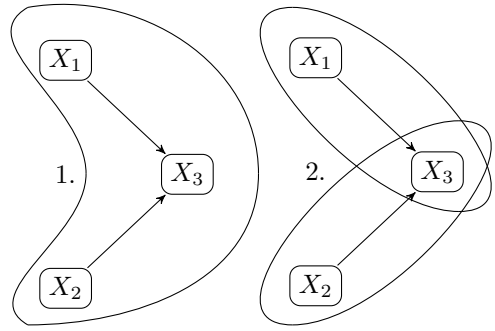


Figure 3: (a) AND gate (b) different systems of dependences

As indicated above, we used a third feature in case of the *Joint GIS* algorithm in order to gain the same limit point as the CGIS algorithm. Figure 2 (b) shows the results of this test. We observe that both algorithms now share the limit point 0 and that CGIS converges considerably faster than *Joint GIS*. The reason for this difference is in this case not the choice of the set X, but the additional feature we introduced for the *Joint GIS* algorithm.

A downside of iterative scaling algorithms is their poor performance compared to gradient methods shown for example by Huang et al. in [11] or by Minka in [12]. That is why we display a third algorithm in Figure 2 (b). This algorithm is a faster version of CGIS, the Sequential Conditional Generalized Iterative Scaling (SCGIS) algorithm presented by Goodman in [10]. Although SCGIS is considerably faster than the other algorithms presented here, it is still not as good as the gradient methods it was compared to by Huang et al. in [11]. This leads to the result that an iterative scaling method may not be the fastest way to calculate a maximum entropy model, but a reliable one.

In conclusion, we were able to fully explain the connection and highlight the differences between the considered Generalized Iterative Scaling algorithms.

References

[1] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information Geometry*. Springer, 2017.

[2] J. B. Brown, P. J. Chase, and A. O. Pittenger. Order independence and factor convergence in iterative scaling. *Linear Algebra and its Applications*, 1993.

[3] S. Chen and R. Rosenfeld. A Gaussian Prior for Smoothing Maximum Entropy Models. Technical report, School of Computer Science, Carnegie Mellon University, 1999.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[5] I. Csiszár. A geometric interpretation of darroch and ratcliff's generalized iterative scaling. *The Annals of Statistics*.

[6] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 1975.

[7] J. R. Curran and S. Clark. Investigating gis and smoothing for maximum entropy taggers. *Proceedings of EACL'03*, 2003. URL: `http://www.aclweb.org/anthology/E/E03/E03-1071.pdf`.

[8] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 1972.

[9] K. Ghazi-Zahedi. Entropy++ github repository. 2017. URL: `http://github.com/kzahedi/entropy`.

[10] J. Goodman. Sequential conditional generalized iterative scaling. *In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2002.

[11] F.-L. Huang, C.-J. Hsieh, K.-W. Chang, and C.-J. Lin. Iterative scaling and coordinate descent methods for maximum entropy models. *Journal of Machine Learning Research*, 2010.

[12] T. P. Minka. Algorithms for maximum-likelihood logistic regression. *Statistics Tech Report*, 2001.

[13] S. D. Pietra, V. D. Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

[14] R. Rosenfeld. *Adaptive Statistical Language Modeling: a Maximum Entropy Approach*. PhD thesis, Carnegie Mellon University, 1992.

[15] D. Yuret. Naive bayes is a joint maximum entropy model. 2010. URL: `http://www.denizyuret.com/2010/11/naive-bayes-is-joint-maximum-entropy.html`.

Proposal of probability risk Evaluation for System Development Project Based on Requirements Analysis and Bayesian estimation

Shinji Mochida
Faculty of Commerce
University of Marketing and Distribution Sciences
mochida@red.umds.ac.jp

Abstract

All companies employ business systems during the development of new software, mechanical equipment or other services. Additionally In most business system development projects are planned in order to reduce amount of running cost or increase benefit. Thus More efficient project management is needed in order to meet schedules and reduce cost. Although there are several stages in business system development projects. There are some factors that cause over cost or schedule delay of projects. Especially finesse in estimating for requirement from customer is the most essential in project managements. Although there are subjective factors to evaluate requirements. Thus evaluating scientifically is needed using requirements analysis and Bayesian estimation in project management.

1 Introduction

Business System development projects are challenging in that are many requirements demanded from customers even while these requirements are proposed with the same priority. Thus it is important to narrow down and prioritize requirements according to their essentiality and criticality to finish on schedule. Although system developers estimate according to the complexity of projects [2], but customers expect the cost to be based on the number of requirements they propose. Then customers and system developers estimate differently, there are often **conflicting estimates**. Thus, this paper proposes cost share rate for business system development projects based on requirements analysis in order to estimate accurately. Cost share rate is defined as the percentage of total cost assigned to each requirement. Cost share rate could distinguish essential requirements. And requirements

that has large cost share rate must have large risk, also should be under strict control. Because big change or modification for essential requirement give large impact to costs or schedule. Then this paper aim to propose a method to identify essential requirements in order to estimate accurately. Then risks are assigned to each requirements with cost share rate and probability. This research show methods in following steps; (see Figure 1). First, this research predict cost for requirement version one of past project. Next this research compares with predicted estimate(version four) and total cost at completion of the past project. Additionally there is essential point estimating is subjective. Thus this paper propose the method how to predict cost share rate accurately using Bayesian estimation.

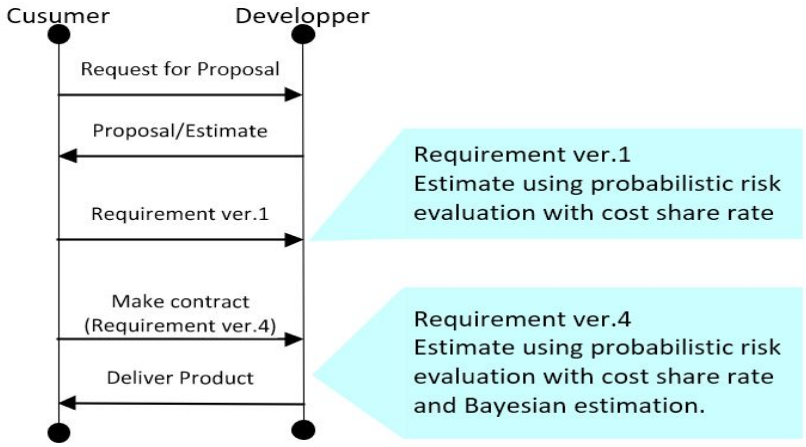


Figure 1: Steps for probabilistic estimation with cost share rate

2 Previous Studies

One purpose of this paper is showing that risk management could contribute reducing cost of business system development project. Additionally this paper demonstrates the potential to evaluate risk by requirements analysis for business system development project management with Bayesian estimation. This research is not focused on estimating costs based on the method to measure amount of source code of system, but rather allocating costs to each requirement. Previous research typically focused on either schedule, cost estimation or productivity. Improving productivity contribute to finish project fine[3]. About changing requirement in mechanical engineering design, one example of requirements analysis research explored the ability of predicting requirements change through graphical models of the requirements documents and historical change trends[4]. There are plus and minus risks in system development project. although there are no research that

refers subjective factor. This propose the method take into subjective factor in account with Bayesian estimation.

3 Project management and Risk management

3.1 Project management and Risk management

In project management there are two important methods[5]. One is schedule management, and another one is risk management. On the other hand Equation 1 shows there are controllable factors or uncontrollable factors in business. Sale is uncontrollable factor, because sale come under the influence of markets, customers. Although cost is controllable factor, cost includes payment, material costs, for example, payment for staff could be cut off by manager. It is important to identify which factor is controllable and uncontrollable. And how appropriately controllable factor could be controlled. Also there is possibility uncontrollable factor could be controlled with Bayesian analyses. Usually extra budget is settled aside for refinement or fixing trouble in project management. This extra budget is called for contingency budget or only contingency. If risk management would work well, contingency budget would not be used, then contingency budget would come to benefit. At the result prospect of profit would increase. Thus risk management could contribute reducing contingency cost,And risk management has potential to increase contingency profit.

$$Gaining = Sale - Cost \tag{1}$$

3.2 Risk analysis

Risk is defined as factors that make uncertain when they will achieve their objectives under ISO31000. Usually risk Analysis is started from risk identification in risk management. Then, risk evaluation is considered by qualitative evaluation and quantitative evaluation. Thus quantitative risk analysis is calculated by possibility $\times$ cost. This calculated risk(cost) should be spend ,if risks comes up. It is called expected monetary value. Risks is evaluated by expected monetary value Equation 2. And risks are prioritized by the order of expected monetary value. In quantitative evaluation for risk management usually probability is given subjectively by staffs subjectively. Or probability is given by experts, Delphi method or questionnaire for skilled staffs. Thus getting accurate probability is very essential to cucullate expected monetary value correctly. Thus this paper takes into account β distribution to calculate probability, and expect monetary value. This paper propose conditional possibility in order to cucullate risk correctly. It is very essential point in risk management there are plus and minus risks, in addition probability is subjective; (see Figure 2).In this research conditional possibility is given to essential

risks. In many project risk comes up by misunderstanding requirements or defects in requirements from clients in business system development project. It is vital to get certain requirements and predict risk in requirements properly.

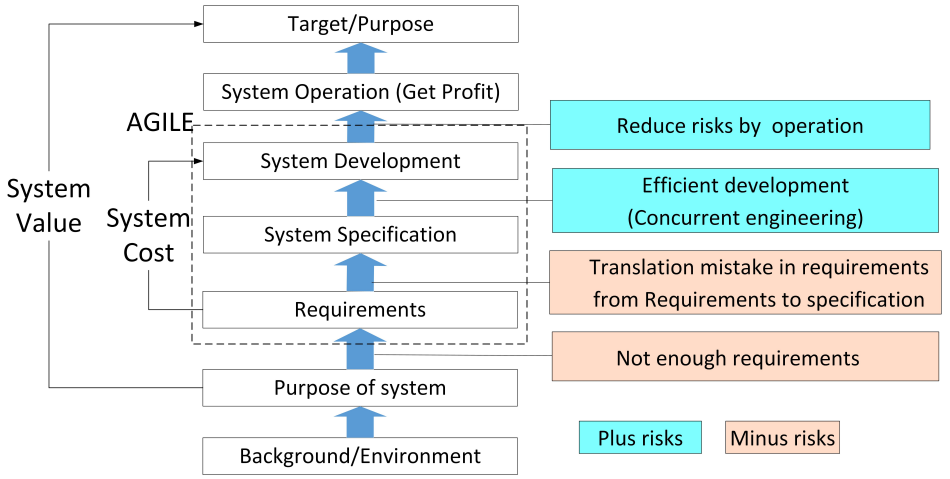


Figure 2: Plus and Minus risks in system development project

$$Risk(Expected\ Monetary\ Cost) = Probability \times Cost \quad (2)$$

4 Cost Prediction Methods for Business System Development

4.1 Cost Prediction Methods

Proper estimate is essential to finish projects on schedule and under budget. Over cost or schedule delay is caused by missing estimate. Usually amount of program source code is predicted by some prediction method in order to estimate in business system development project. Then amount of program source code is converted into base monetary cost. Next total cost is made by adding contingency cost to base monetary cost. It is finally budget for project. Then this section explains some current methods to predict costs of business system development projects. There is typical methods to estimate for business system, typical methods;COCOMO method and Function Point method. Both methods predict costs for business

system. In case of COCOMO method, it estimate by amount of program source code. On the other hand In case of Function Point method it accumulates points according to the complexity of system;the number of db tables, dialog boxes, print forms and interfaces. Acquired points could be converted into cost.

4.2 COCOMO Method

COCOMO Method[6] estimate the duration (Person-Months) in system development projects. In the COCOMO method, volume of source code is estimated by Equation 3. Duration (Person-Months) could be calculated with dividing volume of source code by the number of staffs. COCOMO Method propose Equation 3. And it uses the parameters as follows: C_0,C_e, P_1, P_2, and P_3.

C_e: estimate duration (PM:Person-Months) for expectation
C_0: estimate volume of source code
P_1: parameter for estimated productivity
P_2: exponent parameter for software development
P_3: calibration parameter
A challenge with this method is the parameters used for the cost estimation method are empirically derived and contextually dependent on many different factors, such as team size, project complexity, cultural environment, and others.

$$C_e = (C_0 \times P_1)^{P_2} \times P_3 \tag{3}$$

4.3 Function Point Method

Function Point Method estimate duration (PM:Person-Months) as those of COCOMO methods [5][6]. In the Function point method it is necessary to count the number of internal and external files, tables and internal and external interfaces. Function Point Method propose equation 4.And it uses the parameters as follows: C_e,F_p, F_1, P_1, and P_2.

C_e: estimate duration (PM:Person-Months) for expectation
F_p: estimate function points
F_1: function points
P_1: parameter for estimated productivity
P_2: calibration parameter

$$C_e = (F_p \times P_1), \quad F_p = F_1 \times P_2 \tag{4}$$

4.4 Other Methods to Estimate

Additionally, there is another cost estimate method as Experience method. In the experience method, total cost is estimated based on previous experiences. In these methods, there are often gaps between system developers' cost estimates and customer expectations. This results from differences in how developers and customers group costs. Estimating correctly is important to finish building system on schedule and under budget. Thus, translating requirements into factors to estimate is essential. Certain requirements are needed to estimate properly. But there is no cost estimate method taking certainty of requirements into account. This paper considers taking certainty of requirements into account to estimate in order to gain customer agreement.

5 Requirement Analysis using Linguistic analysis

	Req01	Req02	Req03	Req04	Req05	Req13	Req14	Req15	Req16	Req17	Req18	Req19	Req20	Req21	Req22	Sum
Req01	0	6	3	2	2	1	3	3	1	0	2	4	1	0	0	38
Req02	6	0	3	2	3	1	2	1	1	0	3	3	1	0	0	35
Req03	3	3	0	20	10	0	4	0	1	1	3	8	15	0	1	81
Req04	2	2	20	0	11	0	4	0	0	1	4	8	15	0	1	76
Req05	2	3	10	11	0	1	1	0	0	2	7	16	18	0	1	78
Req06	3	4	8	6	3	0	1	0	1	1	2	2	5	0	0	40
Req07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Req08	1	1	1	1	1	1	0	0	0	0	1	2	2	2	0	18
Req09	2	2	2	0	0	1	1	0	1	0	0	1	1	2	0	19
Req10	4	2	0	0	0	1	1	4	0	0	0	3	0	0	0	15
Req17	0	0	1	1	2	0	0	0	0	0	7	7	1	0	0	20
Req18	2	3	3	4	7	0	1	0	0	7	0	13	5	0	1	50
Req19	4	3	8	8	16	2	1	2	1	7	13	0	10	0	1	84
Req20	1	1	15	15	18	1	1	0	0	1	5	10	0	1	1	79
Req21	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	6
Req22	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0	6
Sum.	38	35	81	76	78	10	22	11	7	20	50	84	79	6	6	702

Figure 3: Count over lapping keywords from requirement version four

Taking correct requirements is essential to estimate properly. And over cost or schedule delay is caused by missing evaluation of requirements. Otherwise over cost or schedule delay is caused by many remediation of requirements. Remediation for essential requirement and uncertain requirement has large risk.thus It is vital to distinguish the requirement which gives large impacts to specification or budget of project. This research propose the method to distinguish influential requirements that has large risks. This paper shows a method to distinguish an influential requirements with with linguistic analysis and cost share rate. Cost share rate is defined as the percentage of total cost assigned to each requirement. This paper analyzes the requirements that were requested in past small system development projects by linguistic analysis. This small project is building a knowledge collecting system. In this project requirements were revised four times. Thus this paper analyzes requirement version one and version four. This paper predict risks from

the result from analyzing requirement version one. Risk is considered as cost in risk management. This results; acquired cost is compared with the actual cost at

	Req01	Req02	Req03	Req20	Req21	Req22
1	system	system	When	following	simultaneou	Schedule
2	achieves	reads	equipment	screen	connected	screen
3	taking	tag	passes	installed	number	schedule
4	out	pasted	gate	as	clients	function
5	equipment	taking	direction	display	PC	
6	efficiency	out	taking	online	assumed	
7	improvemer	worker	out	monitor	50	
8	return	equipment	image	history		
9	managemer	noncontact	vicinity	inspection		
37				assumed		
38				one		
39				renew		
40				still		
41				picture		
42				regularly		

Figure 4: Extract keywords from requirement version four

the completion of project. If risk for requirement could be predicted properly, it would contribute project management. In this research, overlapping keywords are extracted form each requirements with linguistic analysis. Overlapping keywords are words that appear in one requirement and groups of keywords that appear in each category or phase in system development project. Categories or phases are Design, Development, Print, Test, Interface and Document. Overlapping keywords indicates relationships between one requirement and each other. The number of relationships that each requirement has with other requirements indicate essentiality and importance. Steps of linguistic analysis are as follows:
(1) Extract keywords from each requirement in version four. (see Figure 4).

(2) Count overlapping key words from extracted keywords, also count overlapping key words from extracted keywords in version four (see Figure 3).

(3)Distinguish essential requirements by counting the number of over lapping keywords and subjective cost share rate from three engineers, and measure the distance from most essential requirement to each requirement in version four (see Figure 5).

(4)Distinguish essentiality of each requirements from these results, and distinguish categories that each requirements belong in version one.

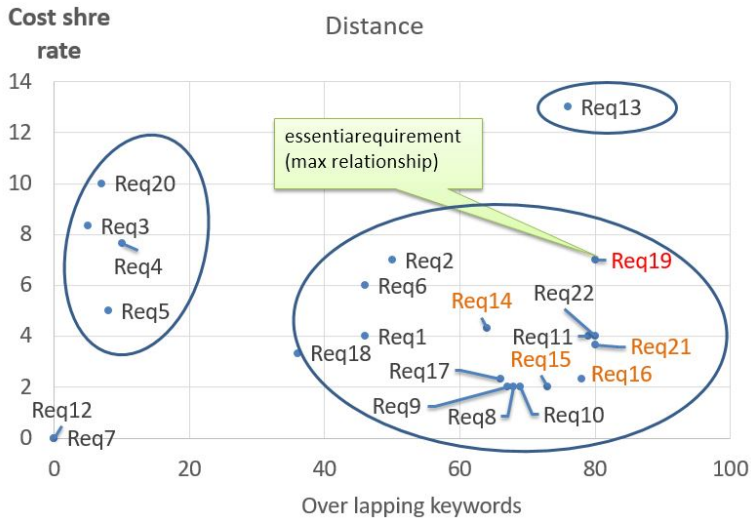


Figure 5: Specify important requirement from requirement version four

6 Prediction of Cost Share Rate

Usually costs for system development projects are estimated by grouping costs with the number of dialog boxes, interfaces or print forms . Alternatively, costs may be estimated by associating cost to logic design, development, test, adjustment and documents; but, not according to the requirements .System developers estimate according to the complexity of projects, but customers expect the cost according to the number of requirements. Thus, customers could not understand the estimates provided by system developers. This paper shows a method to calculate cost share rate for each requirement in order to evaluate requirements accurately with mutual understanding of the developer and customers. Also cost share rate indicates importance of each requirement. Cost share rate is defined as the percentage of total cost assigned to each requirement. Figure 6 sows cost share rate of version our of past project. Although in this research cost rate(%) for each requirements are subjective figure from the two staffs that worked on this system development project. And cost share rates are gained by multiply cost rate by cost under the estimate. This cost rate and estimated cost are the value that according to the category.

7 Probability Prediction of schedule delay

This paper aim to predict risk of system development project based on requirements analysis. This paper consider one of risk of system development project is

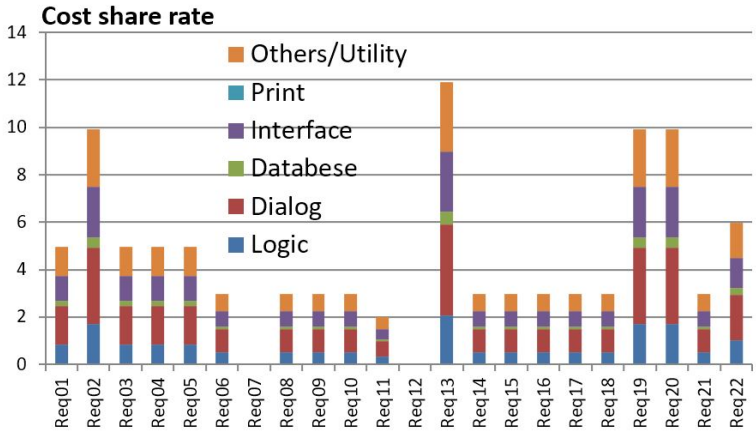


Figure 6: Cost share rate of requirement version four

schedule delay. Thus this research suppose probability of schedule delay follows β distribution (see Equation 5), Sample data in Table 1 are surveyed in past system development project; Rdf system for tool tracking in machine factory. Then parameters (see Table 2) are gained by curve fitting sample data into β distribution. In this analyzing process x is probability parameter that indicates start day for each task, and y is ratio for schedule delay against actual days. Table 1 shows that survey/preparation, design and programming process have risk of schedule delay. Although test and writing document process have no risk of schedule delay. Figure 7 shows β distribution curve in this case from parameters (see Table 2).

$$f(x) = c \times x^{\alpha\ -1}(1-x)^{\beta\ -1} \tag{5}$$

Table 1: Sample data from past project

Item	Survey/	Design			Programing			Test		
(days)	Prepare	Dia.	Fun	DB	Dia.	Fun	DB	Test	Doc.	Sum
Schejule	11	29	27	27	9	58	39	3	13	216
Actual	35	34	36	27	9	55	62	2	7	267
Delay	24	5	9	0	0	-3	23	-1	-6	51
Prog	0	0.52	0.6	0.59	0.7	0.74	0.59	0.95	0.93	

Prog.:Progress Rate Fun:function Doc.:Document Dia.:Dialog

Table 2: Acquire Parameter for β distribution

C	α	β
0.839835	1.020625	3.047617

8 Risk Prediction for business system development project

Figure 7 shows that probability of schedule delay is 0.225. Thus this research

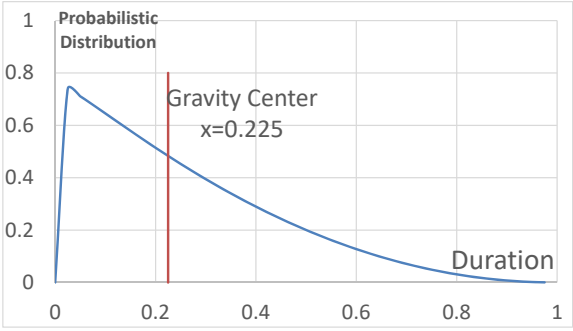


Figure 7: Beta distribution from past project

assign probability of risk;0.225 to essential requirements that distinguished by requirements analysis. Actually Table 3 shows that requirement3 and requirement4 are about design and development, requirement6 is about development. These requirements are essential, and assigned probability of risk as 0.225. Requirement1 is about Main Design, and assigned probability of risk as 0.1,because in past three project contingency is set as10% (see Table 4). Contingency is set for refinement. Refinment contains deleting, adding or refactoring. Table 4 shows occupying cost rate for each work process from surveying of past system development project; medical record system, knowledge management system. Thus risk is calculated probability of risk $\times$ cost (see equation 2),cost is analyzed by cost share rate. In estimated cost of requirement version one estimated total cost is 109 (see Table 3).And at completion actual total cost is 109.1 (see Table 3). At the result these results are equal.

9 Bayesian estimation

First of all, costs is estimated in the project management. Also accuracy of estimation is needed. However, estimate isn't accurate usually.

Table 3: Total cost at completion

Aitem	Attr	Cost Share Rate(%)	Schdule delay Probability	Conditional Probability	Monitary Risk	Estimated Cost
Req1	M	22.5	0	0.1	2.25	24.75
Req2		25	0	0	0	25
Req3	D/De	5	0.225	0	1.125	6.125
Req4	D/De	7.5	0.225	0	1.6875	9.1875
Req5		22.5	0	0	0	22.5
Req6	De	17.5	0.225	0	3.9375	21.4375
Toatal		100			9	109

Req:Requirement M:Main, D:design, De:Development Attr:Attribute

The reason is that there are subjective factors, and the accuracy of the requirements is low. Thus, this paper propose the method to estimate of the project-properly using Bayesian estimation(expression 6) and cost share rate proposed in before section. The example of the adjustment to the cost estimate of past actual projects is shown (see Table 4) This table shows there are over cost as 10% in almost

Table 4: Data from past project

Project	Design	Development	Testing	Document	Refinement	Total
P1	3	4	1	1	1	10
P2	1	6	1	1	1	10
P3	1	6	1	1	1	10
Average	1.7	5.3	1	1	1	10

project. There is subjective factor to estimate, and there is plus and minaus lisks. Also there is inaccuracy in requirements. Figure 8 shows there is valiance of cost share rate according to subjective factors and inaccuracy in requirements .Thus data from past project(Table 5) are inputted into Bayesian estimation(expression 7)[6]. Thus The result was 1.07 is gained. It is important point plus risks are estimated as 1.0. Plus risk are risks that have minus difference. This result shows only minus risks influence cost or estimate. Also this results shows possibility estimate would be gained accurately with Bayesian estimation.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{6}$$

$$P(Obud|Est) = \frac{P(Est|Obud)P(Obud)}{P(Obud)P(Est|Obud) + P(Obud)P(Est|Obud)} \quad (7)$$

$Obud : Proceedon\ budget\ Actually$

$Est : Estimation$

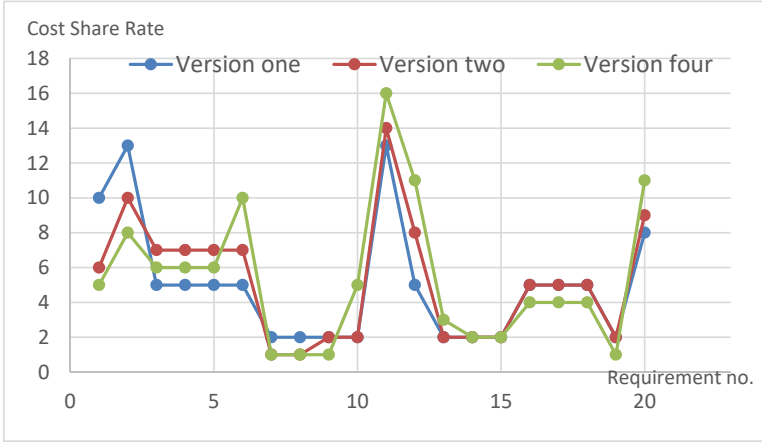


Figure 8: Variation of each cost share rate and revise

10 Conclusion

Meeting budget, finishing on schedule, and maintaining high quality are all important in project management. If cost and duration would be gained accurately, meeting budget and finishing on schedule would be achieved and managing project well. Estimate miss by developer causes finally over cost and schedule delay. Subjective factor and inaccuracy in requirements are reasons why estimate miss is caused. Generally estimate is measured by amount of source code of system or complexity of system. Also it is not estimated based on essentiality or risk of requirements. Misunderstanding of requirements and subjective factors cause misses in estimate or schedule. Then, this paper proposes cost share rate to measure essentiality of requirements in order to estimate accurately. Cost share rate is defined as the percentage of total cost assigned to each requirement Usually on risk management risks are evaluated according to staffs' experience. Also risks are not evaluated based requirement analysis. Evaluating risk properly is needed in order to manage project well. Thus this paper show better results by using cost share

rate from requirement analysis. Additionally, this method helps prioritize requirements and narrow down specifications of the project. Prioritizing requirements and narrowing down specifications accurately help ensure it meets budget and duration targets. Additionally, this paper shows possibility to obtaining probability for risks accurately with cost share rate and Bayesian estimation. However, this result was obtained by small case. Thus further research and study is needed to refine and improve this method to obtain cost share rate and risk more accurately.

Table 5: Bayesian estimation and actual cost by cost share rate

Aitem	First	Middle	Final	Difference	Accuracy	Bayesian estimation	Correction value
Req1	10	6	5	5	0.95	0.99	5.03
Req2	13	10	8	5	0.95	0.99	8.05
Req3	5	7	6	-1	1	1.00	6.04
Req4	5	7	6	-1	1	1.00	6.04
Req5	5	7	6	-1	1	1.00	6.04
Req6	5	7	10	-5	1	1.00	10.06
Req8	2	1	1	1	0.99	1.00	1.01
Req9	2	1	1	1	0.99	1.00	1.01
Req10	2	2	1	1	0.99	1.00	1.01
Req11	2	2	5	-3	1	1.00	5.03
Req13	13	14	16	-3	1	1.00	16.09
Req14	5	8	11	-6	1	1.00	11.06
Req15	2	2	3	-1	1	1.00	3.02
Req16	2	2	2	0	1	1.00	2.01
Req17	2	2	2	0	1	1.00	2.01
Req18	5	5	4	1	0.99	1.00	4.02
Req19	5	5	4	1	0.99	1.00	4.02
Req20	5	5	4	1	0.99	1.00	4.02
Req21	2	2	1	1	0.99	1.00	1.01
Req22	8	9	11	-3	1	1.00	11.06
Total	100	104	107	-7	21.83	21.98	107.63

11 Discussion

Customers and developers estimate costs differently, resulting in differing expectations for project cost. Because in business system development projects there are many ways to implement requirements, there are large variability in translating user's requirement into system specification. It differs greatly according to staff's skill. Additionally subjective factors is another reasons why estimate does not meet final cost. Thus proper estimate by requirement analysis is needed in order to finish business system development project fine. Also correct probability for risk is needed to build proper estimate. Requirements have invisible risks. There are plus risk and minus risk. Minus risk gets prospect of profit worse, but plus risk gets prospect of profit well. But plus risk is not visible, plus risk is only in mind of staffs individually. This is one reason why estimated cost:109 match total cost:109 (see Table 3).Additionally, this paper shows potential accurate estimation would be gained by ruling out plus risks in estimating with Bayesian estimation. Therefore there is potential to predict risk accurately using conditional probability or Bayesian estimation [7][8][9]in order to predict cost accurately.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 17K00354.
The author would like to thank INTERLINK Inc. and TECHNO SOLUTION Inc. for data and advices.

References

[1] Shinji Mochida,James Righter, Joshua D. Summers: *A Study of Project Cost Management Based on Requirements Analysis, International Journal of BMFSA Vol. 21, No. 1* ,jp, 2016. 21-27

[2] Parviz F. Rad, Kou Itou, japanese translation: *PROJECT ESTIMATING AND COST MANAGEMENT* ,Seisansei Syutupan, jp, 2004

[3] Shinji Mochida: *A Study on Method of Measuring Performance for Project Management, 20th ISPE International Conference on Concurrent Engineering* ,IOS Press Ebooks,US, 2013. 264-273

[4] Beshoy Morkos,James Mathieson,Joshua D. Summers: *Comparative analysis of requirements change pre-diction models: manual, linguistic, and neural network, Research in Engineering Design Vol. 25* ,Springer,US, 2014.139-156

[5] Project Management Institute: *A Guide to the Project Management Body of Knowledge Fourth Ed. (JAPANESE TRANSLATION*,Project Management Institute,US, 2008

[6] Kazuya Fujita: *That is Bayes*,OHMSHA,jp, 2015

[7] Christian Robert : *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* ,Springer Texts in Statistics,US, 2007

[8] Paul G Hoel,Akira Asai,Masayasu Murakami japanese translation : *ELEMENTARY STATISTICS* ,Springer Texts in Statistics, jp, 2011

[9] David Blackwell,Kazuyoshi Mikuni japanese translation : *BASIC STATISTICS* ,Kogakusha, jp, 1972

Uncertainty and Stochasticity of Optimal Policies

Guido Montúfar
University of California, Los Angeles
montufar@math.ucla.edu

Johannes Rauh
MPI for Mathematics in the Sciences
jrauh@mis.mpg.de

Nihat Ay
MPI for Mathematics in the Sciences
Santa Fe Institute
University of Leipzig
nay@mis.mpg.de

Abstract

We are interested in action selection mechanisms, policies, that maximize an expected long term reward. In general, the identity of an optimal policy will depend on the specifics of the problem, including perception and memory limitations of the agent, the system's dynamics, and the reward signal. We discuss results that allow us to use partial descriptions of the observations, state transitions, and reward signal, in order to localize optimal policies to within a subset of all possible policies. These results imply that we can reduce the search space for optimal policies, for all problems that share the same general properties. Moreover, in certain cases of interest, we can identify the policies that produce the same behaviors and the same expected long term rewards, thereby further reducing the search space.

1 Introduction

We study stochasticity of optimal policies in decision making. We want to understand under which conditions a deterministic behaviour is optimal.

In many contexts, optimal policies are deterministic: in each situation, there is an action (not necessarily unique) that can be considered as the optimal action in this situation. Thus, an optimal policy for a decision maker can be implemented algorithmically as a mapping from situations to actions. Examples are Markov Decision Problems (MDPs). In an MDP the variables determining the immediate reward are available to the decision maker. One can show that in an MDP, there always exists a deterministic optimal policy.

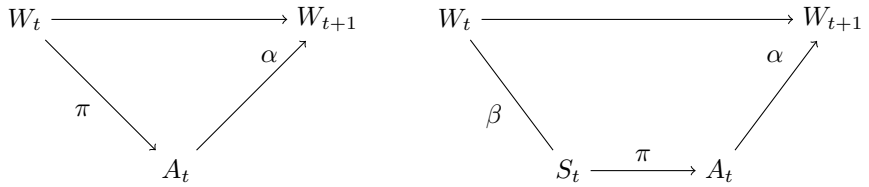


Figure 1: The graphical structure of an MDP and a POMDP.

However, there are situations where optimal policies are not deterministic. The classical example is game theory, as illustrated by the rock-paper-scissors game: when the game is repeated several times, any fixed deterministic strategy can be learned by the opponent, who will then win the game.

Another example is given by Partially Observed Markov Decision Processes (POMDPs). In this case, only a stochastic measurement of the relevant variables is available to the decision maker. In order to be more precise, we now introduce the basic setting and the corresponding notation (see also the graphical representation in Figure 1):

- W_t – world state
- S_t – sensor value
- A_t – chosen action at time t
- α, β – fixed transitions
- π – policy ($\rightarrow$optimize)
- After each step, agent receives reward $R(w_t, a_t)$.

What are the mechanisms that lead to stochastic optimal policies? The game theoretic setting and POMDPs share the following properties:

- **Uncertainty:** The reward R depends on unobserved quantities:
  1. the opponent's strategy
  2. the world state W_t
- **Feedback:** Actions influence the hidden state:
  1. The opponent observes my strategy and adapts.
  2. A_t influence W_{t+1} via α.

Can we control policy stochasticity by controlling uncertainty? One result in this direction is due to [4] (which generalizes [2]). We formulate this kind of problems as localization of optimal policies, and formulate various scenarios in Section 4.

2 Definitions

We consider a POMDP defined by a tuple $(W, S, A, \alpha, \beta, R)$, where W, S, A are finite sets of world states, sensor states, and actions, $\beta\colon W \to \Delta_S$ and $\alpha\colon W \times$

$A \to \Delta_W$ are Markov kernels describing sensor measurements and world state transitions, and $R\colon W \times A \times W \to \mathbb{R}$ is a reward signal depending on the current world state, chosen action, and resulting world state. It is also useful to consider the expected value over resulting world states, $R\colon W \times A \to \mathbb{R}$, with $R(w,a) = \sum_{w' \in W} \alpha(w'|w,a) R(w,a,w')$. A policy is a mechanism for selecting actions. We will focus on stationary (memoryless and time independent) policies, which we simply call *policies*, described by Markov kernels of the form $\pi\colon S \to \Delta_A$. We denote the set of policies by $\Delta_{S,A}$. The deterministic policies map each sensor state to one specific action. They correspond to the vertices of $\Delta_{S,A}$.

The world state is updated at discrete time step by iterating the kernels β, π, α. The objective of learning is to find a policy that maximizes some form of expected long term reward. We focus on the average reward, which for an initial world state distribution $\mu \in \Delta_W$ and a policy $\pi \in \Delta_{S,A}$, is given by

$$\mathcal{R}_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}_{\pi,\mu}\left[\frac{1}{T}\sum_{t=0}^{T-1} R(W_t, A_t, W_{t+1})\right]. \tag{1}$$

We will make the standard assumption that for each fixed policy, the Markov chain of world states is irreducible and aperiodic. This implies that there is a unique stationary limit distribution $p^\pi \in \Delta_W$ of world states. Moreover, this limit distribution is independent of the initial distribution μ. In this case the average reward can be written as

$$\mathcal{R}(\pi) = \sum_w p^\pi(w) \sum_a \sum_s \pi(a|s)\beta(s|w) \sum_{w'} R(w,a,w')\alpha(w'|w,a). \tag{2}$$

Although we restrict the exposition to average rewards as defined above, we note that some of the results hold as well in the setting of discounted rewards, where the rewards in (1) are not weighted uniformly but by a factor γ^t with $\gamma \in (0,1)$.

3 The optimization problem

Before we present the main localization results of this paper, we first explore the structure of the optimization problem in terms of simple examples. They highlight the geometry that underlies the stochasticity of optimal policies.

The objective function (2) reduces to

$$\mathcal{R}(\pi) = \sum_{w,a} p^\pi(w) p^\pi(a|w) R(w,a), \tag{3}$$

if we introduce the effective state policy $p^\pi(a|w) = \sum_s \beta(s|w)\pi(a|s)$. The stationary state distribution $p^\pi(w)$ is the solution of

$$(T^{(\alpha,\beta,\pi)} - I)p = 0 \quad \text{and} \quad \sum_w p_w = 1 \quad \text{and} \quad p_w \geq 0, \tag{4}$$

where we have defined the state transition matrix

$$T^{(\alpha,\beta,\pi)} = [p^\pi(w'|w)]_{w',w} = \left[\sum_a \alpha(w'|w,a)p^\pi(a|w)\right]_{w',w}. \tag{5}$$

In the following example we give (3) in an explicit form, involving only algebraic operations on π (note that we also use the somewhat more suggestive index notation, for instance $\pi_{1|2}$ instead of $\pi(1|2)$). This gives us a sense of the structure of the optimization problem.

Example 1 (Optimization for two states, two actions, two observations)**.** Let $W = \{1,2\}$, $S = \{1,2\}$, $A = \{1,2\}$. In this case we have

$$\mathcal{R}(\pi) = ((R_{11} - R_{12})\xi_{1|1} + R_{12})p_1 + ((R_{21} - R_{22})\xi_{1|2} + R_{22})(1 - p_1), \tag{6}$$

where the effective world state policy

$$\xi_{1|1} = \beta_{1|1}\pi_{1|1} + (1 - \beta_{1|1})\pi_{1|2} \quad \text{and} \quad \xi_{1|2} = \beta_{1|2}\pi_{1|1} + (1 - \beta_{1|2})\pi_{1|2}, \tag{7}$$

and the stationary state distribution obtained by solving (4)

$$p_1 = \frac{(\alpha_{1|21} - \alpha_{1|22})\xi_{1|2} + \alpha_{1|22}}{(\alpha_{1|21} - \alpha_{1|22})\xi_{1|2} + \alpha_{1|22} + (\alpha_{2|11} - \alpha_{2|12})\xi_{1|1} + \alpha_{2|12}}. \tag{8}$$

Note that the denominator can only vanish if the nominator also vanishes.

So, the objective function $\mathcal{R}(\pi)$ is a rational function of degree 2 in $(\pi_{1|1}, \pi_{1|2}) \in [0,1]^2$, with coefficients depending on α, β, and R. We can plot $\mathcal{R}$ over $\Delta_{S,A} \cong [0,1]^2$. Examples are shown in Fig. 2.

Example 2 (Optimization for two states, two actions, blind agent)**.** Let $W = \{1,2\}$, $S = \{1\}$, $A = \{1,2\}$. We set $\beta(s = 1|w = 1) = 1$, $\beta(s = 1|w = 2) = 1$, and $\pi(a|s) = \pi(a)$, in eq. (6), and solve $\nabla_\pi \mathcal{R}(\pi) = 0$. Using the symbolic mathematics library `SymPy`, we find

$$\pi_1 = \frac{-(\alpha_{122} + \alpha_{212})C \pm \sqrt{(\alpha_{121}\alpha_{212} - \alpha_{122}\alpha_{211})CD}}{C(\alpha_{121} - \alpha_{122} + \alpha_{211} - \alpha_{212})}, \tag{9}$$

where

$$C = ((R_{11} - R_{12})(\alpha_{121} - \alpha_{122}) + (R_{21} - R_{22})(\alpha_{211} - \alpha_{212})) \tag{10}$$
$$D = ((R_{11} - R_{21})(\alpha_{122} + \alpha_{212}) - (R_{12} - R_{22})(\alpha_{121} + \alpha_{211})). \tag{11}$$

These are critical points of the objective function, which might be negative or larger than 1.

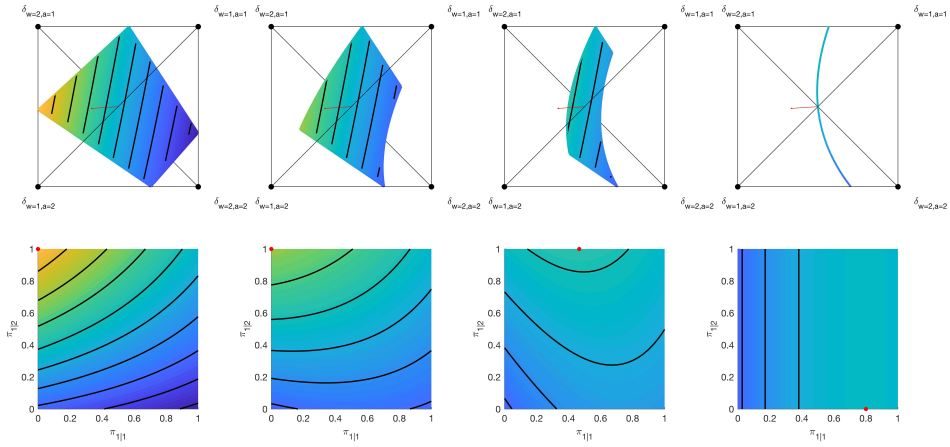


Figure 2: A random choice of α, R, and four choices of $\beta_{s|w}$ going from $[0,1;1,0]$ (fully observable) to $[1,1;0,0]$ (blind). Top row shows the joint distributions $p^\pi(w,a)$. Bottom row shows the policies $(\pi_{1|1},\pi_{1|2})$. Color codes the expected reward $\mathcal{R}(\pi)=\sum_{w,a}p^\pi(w,a)R_{w,a}$. The reward is linear in the space of joint distributions over w and a. The kernel α defines a slice in that space, and β certain inequality constraints. Note that, even in the fully observable case, the expected reward is not a linear function of the policy.

4 Localization of optimal policies

We are interested in saying something about the location of optimal policies. Let $\mathcal{M}$ be a subset of all possible policies. Let $\boldsymbol{A}$ be a subset of all possible state transition kernels $W\times A\to\Delta_W$. Let $\boldsymbol{B}$ be a subset of all possible observation kernels $W\to\Delta_S$. Let $\boldsymbol{R}$ be a subset of all possible reward functions $W\times A\to\mathbb{R}$.

Problem 3. Given $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{R}$, find a subset $\mathcal{N}=\mathcal{N}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{R},\mathcal{M})\subseteq\mathcal{M}$ such that, for any POMDP (W,S,A,α,β,R) with $\alpha\in\boldsymbol{A}$, $\beta\in\boldsymbol{B}$, $R\in\boldsymbol{R}$, there is a policy $\pi^*\in\mathcal{N}$ that is optimal among all policies in $\mathcal{M}$. Ideally, the set $\mathcal{N}$ should be minimal.

We call $\mathcal{N}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{R},\mathcal{M})$ a (*minimal*) *solution set* over $\mathcal{M}$ for the POMDP class $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{R})$. In this language, Theorem 4 below shows that $\mathcal{N}=\{\pi\in\Delta_{S,A}\colon|\operatorname{supp}(\pi(\cdot|s))|\leq k_s\}$ is a minimal solution set over $\mathcal{M}=\Delta_{S,A}$, for the POMDPs with observation kernels from $\boldsymbol{B}=\{\beta\in\Delta_{W,S}\colon|\operatorname{supp}(\beta(s|\cdot))|\leq k_s\}$.

Knowing that there is a set $\mathcal{N}\subseteq\Delta_{S,A}$ which contains an optimal policy allows us to focus the search for optimal policies to the set $\mathcal{N}$. As proposed in [1, 2], this can be used to define a suitable policy model with a reduced number of parameters, without compromising our ability to maximize the average reward.

Observation model

When the world state can be fully recovered from the sensor value, there is an optimal policy within the set of deterministic policies (see, e.g., [5]). This holds irrespective of the specific reward signal and the world state transition kernel. This is an important and well known result in the theory of Markov decision processes. On the other hand, when the agent is blind, it is not possible to localize an optimal policy without taking other specific properties of the system into account.

The following Theorem 4 is a result from [4] refining results from [2]. It generalizes the above discussion to cases where the agent can partially recover the underlying world state. If an observation identifies the world state, there is an optimal policy that is deterministic on this observation. More generally, if a sensor state can result from at most k world states, then there is an optimal policy that randomizes at most k actions at this sensor state. This holds irrespective of the specific reward R and the world state transition kernel α.

Theorem 4. *Consider a POMDP $(W, S, A, \alpha, \beta, R)$. Then there is a policy $\pi^* \in \Delta_{S,A}$ with $|\operatorname{supp}(\pi^*(\cdot|s))| \leq |\operatorname{supp}(\beta(s|\cdot))|$ for all $s \in S$, and $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ for all $\pi \in \Delta_{S,A}$. Moreover, there are POMDPs $(W, S, A, \alpha, \beta, R)$ where each policy $\pi^* \in \Delta_{S,A}$ that is optimal among all policies satisfies $|\operatorname{supp}(\pi(\cdot|s))| \geq |\operatorname{supp}(\beta(s|\cdot))|$.*

One might think that the randomization of actions in a POMDP simply allows the agent to assign weights to the optimal deterministic actions that he would choose if he knew the underlying world state. However, the situation is more subtle: being uncertain about the underlying world state, the agent might need to take distance from actions that have the potential of causing a catastrophic outcome when performed in the wrong state, causing him to choose totally different, more conservative, actions.

Transition model

It is also interesting whether we can localize optimal policies given some information about the world state transition kernel α. For instance, [6] studied blind policies for POMDPs and showed that, if each kernel $\alpha(\cdot|\cdot, a)\colon W \to \Delta_W$, $a \in A$, is symmetric and the reward R is proportional to the starting distribution, then there exists a deterministic optimal blind policy. This result can be slightly generalized as follows, dropping the condition on the reward signal.

Theorem 5. *If all kernels $\alpha(\cdot|\cdot, a)\colon W \to \Delta_W$, $a \in A$, have the same stationary distribution and $\beta(\cdot|w) \in \Delta_S$ is independent of $w \in W$, then there is a deterministic policy $\pi^* \in \Delta_{S,A}$ with $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ for all $\pi \in \Delta_{S,A}$.*

As an example for the first assumption one can consider doubly stochastic matrices, which all have the uniform distribution as stationary distribution. The second assumption means for practical purposes that the agent is blind.

Let us briefly discuss this result in relation to Theorem 4. Assuming that the sensor kernel β always outputs the same sensor state s, the bound of Theorem 4 is

trivial since it upper bounds the number of actions by $\min\{|W|, |A|\}$. Nonetheless, under the above assumptions, Theorem 5 guarantees the existence of an optimal policy that is deterministic. This shows that incorporating properties of the transitions α, in addition to the observation kernels β, allows us to localize optimal policies even further. The next Theorem 6 extends this line of reasoning by taking some of the structure of the reward R into account.

Reward signal

We consider the case where the reward signal takes the form $R(w, a, w') = R(w, w')$, that is, depending on the sequences of world states but not on the specific actions taken by the agent. For instance, one may be interested in rewarding the motion of a robotic arm, without regard of the specific torques applied to the articulations in order to obtain this motion.

For this type of reward signal, the system can be studied in terms of the world state transitions. Each policy π corresponds to a world state transition $p^\pi \colon W \to \Delta_W$ with $p^\pi(w'|w) = \sum_a \sum_s \beta(s|w)\pi(a|s)\alpha(w'|w, a)$. Since the reward only depends on these transitions, any two policies that represent the same world state transition can be regarded as being equivalent. This allows us to restrict the search for optimal policies to a set of unique representatives.

In [3] it is shown that any feasible world state transition kernel can be represented in terms of a policy π with $|\operatorname{supp}(\pi)| \leq |S| + d_{\alpha,\beta}$. Here $d_{\alpha,\beta}$ is the dimension spanned by the vectors $(\beta(s|w)(\alpha(w'|w, a_0) - \alpha(w'|w, a)))_{w \in W, w' \in W} \in \mathbb{R}^{W \times W}$, for $s \in S$, $a \in A \setminus \{a_0\}$, for any fixed $a_0 \in A$. This is the rank of the linear map from policies to world state transition kernels. In particular, there is an optimal policy that satisfies this property, which implies the following result.

Theorem 6. *Consider a POMDP $(W, S, A, \alpha, \beta, R)$ with $R(w, a, w') = R(w, w')$. Then there is a policy $\pi^* \in \Delta_{S,A}$ with $|\operatorname{supp}(\pi)| \leq |S| + d_{\alpha,\beta}$ and $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ for all $\pi \in \Delta_{S,A}$.*

Let us comment this result in intuitive terms. It states that if the reward R is not sensitive to the inner working of the control, such that $R(w, a, w')$ is not dependent on a but only on the outcome w' as result of a, then a number of actions can be ignored when maximizing the expected long term reward. The required number of actions essentially involves the dimensionality $d_{\alpha,\beta}$ of the effect that the control has in the world. Note that this number involves α and β in an entangled way.

5 Conclusion

Given a set of policies and a description of the system, we searched for a smallest set of policies that is guaranteed to contain an optimal policy. In particular, we have discussed how the randomization that is needed in optimal actions is related to the amount of information available to an agent at the moment of deciding on the actions. The results presented here can be summarized as follows:

- If the world state is observable, then there is an optimal policy that is deterministic.
- If an observation uniquely identifies the world state, then there is an optimal policy that is deterministic on this observation.
- More generally, if an observation results from at most k world states, then there is an optimal policy which randomizes at most k actions on this observation.
- If all world transition kernels (indexed by the actions) have the same stationary distribution and the observation kernel is independent of the world state, then there is an optimal policy that is deterministic.
- If the reward signal depends on the current and the future world states but not on the specific actions taken, then there is an optimal policy within a low-dimensional face of the set of all possible policies.

References

[1] N. Ay, G. Montúfar, and J. Rauh. Selection criteria for neuromanifolds of stochastic dynamics. In Y. Yamaguchi, editor, *Advances in Cognitive Neurodynamics (III)*, pages 147–154. Springer, 2013.

[2] G. Montúfar, K. Ghazi-Zahedi, and N. Ay. Geometry and determinism of optimal stationary control in POMDPs. *arXiv:1503.07206*, 2015.

[3] G. Montúfar, K. Ghazi-Zahedi, and N. Ay. A theory of cheap control in embodied systems. *PLoS Computational Biology*, 11(9):1–22, 2015.

[4] G. Montúfar and J. Rauh. Geometry of policy improvement. In *Geometric Science of Information, LNCS 10589*, pages 282–290. Springer, 2017.

[5] S. M. Ross. *Introduction to Stochastic Dynamic Programming: Probability and Mathematical*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, Inc., 1983.

[6] N. Vlassis, M. L. Littman, and D. Barber. On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory*, 4(4):12:1–12:8, 2012.

Probabilistic interpretations of argumentative attacks: logical and experimental foundations*

Niki Pfeifer
Munich Center for Math. Phil.
LMU Munich
niki.pfeifer@lmu.de

Christian G. Fermüller
Dept. for Logic & Computation
TU Vienna
chrisf@logic.at

Abstract

We present an interdisciplinary approach to study systematic relations between logical form and attacks between claims in an argumentative framework. We propose to generalize qualitative attack principles by quantitative ones. Specifically, we use coherent conditional probabilities to evaluate the rationality of principles which govern the strength of argumentative attacks. Finally, we present an experiment which explores the psychological plausibility of selected attack principles.

1 Introduction

Various disciplines study argumentation, including computer science (e.g., [6, 1]), philosophy (e.g., [21]), and psychology (e.g., [11, 13]). Our approach is an interdisciplinary one, as we combine elements of Dung-style abstract argumentation [6], logical argument forms, coherent conditional probability, and also present an experimental assessing the descriptive validity of selected formal principles.

We investigate systematic relations between logical form and attacks between claims in an argumentative framework. Argumentation is a highly complex and dynamic process. Usually, arguments are conceived as premise ("support") and conclusion ("claim") pairs. We focus on static argumentation and are only interested in claims formalized by classical propositional formulæ.

The outline of the paper is as follows: Section 2 gives a brief survey of qualitative attack principles which were investigated in a modal logical framework [3]. We argue, that the modal logical framework appears to be too coarse, especially for

*We thank Gernot Salzer for making the experiment possible during his class: thanks also to his students for their participation. Niki Pfeifer is supported by his DFG project PF 740/2-2 (part of the SPP1516).

modelling the quantitative dimension of attack principles. We therefore propose to generalize these principles by adopting a probabilistic framework. Specifically, we use coherent conditional probabilities to evaluate systematically the rationality of attack principles: coherence provides a criterion for selecting attach principles (i.e., "good" principles should be coherent). In Section 3 we show how to model the qualitative attack principles in probabilistic terms. Section 4 presents our probabilistic analysis of the quantitative attack principles and their semantics. Section 5 presents an experiment which aims to explore the psychological plausibility of selected quantitative attack principles. Section 6 concludes the paper by some remarks on future research.

2 Qualitative attack principles

In what follows we write "$A \longrightarrow B$" to denote that there is an argument claiming A that attacks an argument with claim B. Thus, strictly speaking, attack relations are between arguments. However, we simply say "A attacks B".

It seems intuitively obvious that given attacks on claims implicitly entail attacks on further claims which logically imply the original, explicitly attacked claims. A corresponding 'general attack principle' has been formulated in [3]:

(A.gen) If $F \longrightarrow A$ and $B \models A$ then $F \longrightarrow B$.

While it may be problematic to consider *all* classical logical implicants as inducing implicit attacks in this manner, at least the following instances of **(A.gen)** seem reasonable, since they are immediate and hold even if the consequence relation ($\models$) is constrained to minimal logic [12].

(A.$\wedge$) If $F \longrightarrow A$ or $F \longrightarrow B$ then $F \longrightarrow A \wedge B$.

(A.$\vee$) If $F \longrightarrow A \vee B$ then $F \longrightarrow A$ and $F \longrightarrow B$.

(A.$\supset$) If $F \longrightarrow A \supset B$ then $F \longrightarrow B$.

Actually, **(A.$\supset$)** may raise concerns, since $A \supset B$ does not *relevantly* follow from B, cf. [7]. Hence, one may prefer the following weaker rationality postulate, instead.

(B.$\supset$) If $F \longrightarrow B$ and $F \not\longrightarrow A$[1] then $F \longrightarrow A \supset B$.

Concerning negation, the following principle is intuitively plausible.

(A.$\neg$) If $F \longrightarrow A$ then $F \not\longrightarrow \neg A$.

On the other hand, one can formulate inverse forms of the above principles:

(C.$\wedge$) If $F \longrightarrow A \wedge B$ then $F \longrightarrow A$ or $F \longrightarrow B$.

(C.$\vee$) If $F \longrightarrow A$ and $F \longrightarrow B$ then $F \longrightarrow A \vee B$.

[1] $F \not\longrightarrow A$ denotes that A is *not* attacked by F.

(C.⊃) If $F \longrightarrow A \supset B$ then $F \longrightarrow B$ and $F \not\longrightarrow A$.

(C.¬) If $F \not\longrightarrow A$ then $F \longrightarrow \neg A$.

These last mentioned principles seem, at least partly, to be intuitively much more demanding than those following from **(A.gen)**. Indeed, the results of [3] imply that imposing *all* of the above (connective specific) attack principles amounts to an alternative characterization of classical logic, while proper subsets of the full set of these principles lead to weaker logics that result from discarding some of the logical inference rules of Gentzen's classical sequent calculus **LK**.

The indicated situation calls for a robust interpretation of the attack relation that is capable of formally supporting (or questioning, as appropriate) informal intuitions about the varying strength of the attack principles. To this aim the authors of [3] suggest to translate $F \longrightarrow A$ into the modal formula $\Box(F \wedge \neg G)$, where in the underlying Kripke frame $\langle W, R\rangle$, W models the set of possible states of affairs and wRv is read as "v is a possible alternative from the viewpoint of w". In other words, this setup suggests that a given attack refers to all possible states of affairs in which the attacking claim holds and asserts that the attacked claim does not hold in any of those states. If one stipulates that R is reflexive (or at least serial) than this interpretation of $F \longrightarrow A$ renders the principles **(A.∧)**, **(A.∨)**, **(C.∨)**, **(C.⊃)**, and **(A.¬)** formally sound while one may construct counter examples for the translations of the principles **(C.∧)**, **(C.¬)**, **(B.⊃)**, and therefore also of **(A.⊃)**. Since this result is unsatisfying, in particular with respect to the arguably counter-intuitive classification of attack principles for implication, three alternative modal interpretations where briefly discussed in [3] as well. However, each of the suggested translations of $F \longrightarrow A$ into modal logic is too coarse, since there seems be no principled way to disentangle strong and weak attack principles. Moreover, modal logic does not support quantitative refinements of the qualitative attack principles.

3 Probabilistic semantics

In light of the results of [3], as sketched in Section 2, the challenge to come up with an intuitively convincing and formally sound interpretation of the attack relation between claims of arguments remains open. This motivates us to explore to which extent one may employ coherence-based *conditional probability* (see, e.g., [2, 9]) for this purpose. Concretely, we suggest to read "F attacks A" as the assertion that it is likely that A does not hold, given that F holds. More precisely, we interpret $F \longrightarrow A$ by $p(\neg A|F) \geq t$ for some threshold $0.5 < t \leq 1$. Throughout the paper, we assume that F is not a logical contradiction (i.e., F is not equivalent to $\bot$), since otherwise the corresponding conditional probability is undefined.

Translating the attack principles that refer to conjunction, disjunction, and negation according to the suggested interpretation is straightforward. The following claims correspond to the 'weak' principles **(A.∧)**, **(A.∨)**, and **(A.¬)**:

$(\mathbf{A.}\wedge)_p$ If $p(\neg A|F) \geq t$ or $p(\neg B|F) \geq t$, then $p(\neg(A \wedge B)|F) \geq t$.

$(\mathbf{A.}\vee)_p$ If $p(\neg(A \vee B)|F) \geq t$, then $p(\neg A|F) \geq t$ and $p(\neg B|F) \geq t$.

$(\mathbf{B.}\neg)_p$ If $p(\neg A|F) \geq t$, then $p(\neg\neg A|F) = p(A|F) < t$.

Analogously, the inverse ('strong') principles translate as follows:

$(\mathbf{C.}\wedge)_p$ If $p(\neg(A \wedge B)|F) \geq t$ then $p(\neg A|F) \geq t$ or $p(\neg B|F) \geq t$.

$(\mathbf{C.}\vee)_p$ If $p(\neg A|F) \geq t$ and $p(\neg B|F) \geq t$ then $p(\neg(A \vee B)|F) \geq t$.

$(\mathbf{C.}\neg)_p$ If $p(\neg A|F) < t$ then $p(\neg\neg A|F) = p(A|F) \geq t$.

It is straightforward to check the following.

Proposition 1. $(\mathbf{A.}\wedge)_p$, $(\mathbf{A.}\vee)_p$, $(\mathbf{B.}\neg)_p$, *and* $(\mathbf{C.}\neg)_p$ *hold in the sense of coherence-based probability logic. However,* $(\mathbf{C.}\wedge)_p$ *and* $(\mathbf{C.}\vee)_p$ *do not hold in this sense.*

Note that our probabilistic interpretation of the attack relation, justifies not only **(A.¬)**, but also the intuitively more demanding principle $(\mathbf{C.}\neg)_p$. This is a consequence of the fact that we insist on classical negation here and hence have $p(\neg\neg A) = p(A) = 1 - p(A)$. It might be worth mentioning that actually both $(\mathbf{B.}\neg)_p$ and $(\mathbf{C.}\neg)_p$ cease to hold if one admits .5 as a threshold value. Another interesting observation is that for $t = 1$ **(C.∨)** is justified, since: if $p(\neg A|F) = 1$ and $p(\neg B|F) = 1$, then $p(\neg(A \vee B)|F) = p(\neg A \wedge \neg B|F) = 1$ is coherent (cf. the probabilistic version of the And rule of System P, [9]).

Interpreting attack principles involving the implication connective is more delicate, since it is widely agreed that the natural language conditional ('if ... then ...') should not be identified with classical (truth-functional) implication. Actually, as argued, e.g., in [10, 15], coherence-based conditional probability itself provides a sound and robust semantics for the conditional. Following this insight would force us to use degrees of beliefs in nested conditionals (e.g., in terms of previsions in conditional random quantities; see, e.g., [19, 20]) to interpret principles like **(A.⊃)**. While this is an interesting topic for future research, here we only want to check how our probability-based interpretation of the attack relation classifies **(B.⊃)**, **(A.⊃)**, and **(C.⊃)**, if we replace $A \supset B$ by $\neg A \vee B$. The corresponding translations are as follows:

$(\mathbf{A.}\supset)_p$ If $p(\neg B|F) \geq t$ then $p(\neg(A \supset B)|F) \geq t$.

$(\mathbf{B.}\supset)_p$ If $p(\neg B|F) \geq t$ and $p(\neg A|F) < t$, then $p(\neg(A \supset B)|F) \geq t$.

$(\mathbf{C.}\supset)_p$ If $p(\neg(A \supset B)|F) \geq t$ then $p(\neg B|F) \geq t$.

$A \supset B = \neg A \vee B$ turns $(\mathbf{A.}\supset)_p$ and $(\mathbf{C.}\supset)_p$ into instances of $(\mathbf{A.}\vee)_p$ and $(\mathbf{C.}\vee)_p$, respectively. Moreover, $(\mathbf{B.}\supset)_p$ follows from $(\mathbf{A.}\supset)_p$. Consequently we obtain:

Proposition 2. $(\mathbf{A.}\supset)_p$ *and* $(\mathbf{B.}\supset)_p$ *both hold in the sense of coherence-based probability logic, but* $(\mathbf{C.}\supset)_p$ *does not hold in this sense.*

In [3] also logically contradictory claims are considered by formulating the following corresponding attack principle:

(A.⊥) For every F: $F \longrightarrow \bot$.

In other words, it is stipulated that every argument (implicitly or explicitly) attacks contradictory claims. We may observe that this assumption is in line with our interpretation of the attack relation, since $p(\neg\bot|F) = 1$. However, note that we cannot interpret any principles that involve contradictory claims of attacking arguments, since the corresponding conditional probability must remain undefined.

4 Quantitative attack principles & their semantics

So far, we have only discussed qualitative attack principles, i.e., principles that only care for the presence or absence of an attack between (claims of) given arguments. However it is natural to refine such an analysis by considering *weights* or *varying strength* of attacks. Various suggestions regarding so-called weighted argumentation frames can be found in the literature on argumentation in AI, see, e.g., [8, 5]. But, similarly to the qualitative scenario, there is as yet hardly any analysis of rationality postulates that systematically relates weights of explicit and implicit attacks to the *logical form* of involved claims of arguments. A first step in that direction has been attempted in [4], where the principles introduced in [3] are generalized to the context of weighted argumentation frames. The aim of [4] is to explore under which assumptions one can characterize various t-norm based fuzzy logics in terms of 'weighted attack principles'. As expected, it turns out that some of the principles that are needed to recover a truth-functional (fuzzy) semantics are implausible from an intuitive, argumentation based point of view. In any case, the situation, once more, calls for a systematic interpretation of the relevant principles, that enables one to formally judge their respective plausibility.

Rather than just distinguishing between $F \longrightarrow A$ and $F \not\longrightarrow A$ ("F attacks / does not attack A"), we will use $F \xrightarrow{w} A$ to denote that F attacks A with weight (or degree) w. The corresponding weights are understood to be normalized, with 1 being the maximal weight of any attack, whereas $F \xrightarrow{0} A$ means that F in fact does not attack the claim A at all. Note that this stipulation entails that the qualitative scenario discussed in sections 2 and 3 amounts to an instance of the weighted case, where the only possible weights are 0 and 1.

An attractive feature of the probabilistic approach taken here is the fact that it immediately leads to a quantitative refinement of the qualitative case: interpreting attacks in terms of coherent conditional probabilities suggests to directly attach weights, instead of using thresholds to judge whether a given statement attacks another one. As pointed out in [4], there are several non-equivalent ways in which the the qualitative attack principles reviewed in Section 2 can be generalized to 'weighted attack principles'. The most straightforward generalization of principle **(A.∧)** to weighted attacks is arguably the following:

$(\mathbf{A}^w.\wedge)$ If $F \overset{x}{\longrightarrow} A$ and $F \overset{y}{\longrightarrow} B$, then $F \overset{z}{\longrightarrow} A \wedge B$, where $z \geq \max(x, y)$.

Actually, since we also consider attacks of weight 0 (interpreted as 'no attack'), we may assume without loss of generality that there is a weighted attack between any pair of formulæ. This means that $(\mathbf{A}^w.\wedge)$ can be reformulated as a constraint on the corresponding weights, s.t.:

$(\mathbf{G}^w_{\geq}.\wedge)$ If $F \overset{x}{\longrightarrow} A$, $F \overset{y}{\longrightarrow} B$, and $F \overset{z}{\longrightarrow} A \wedge B$, then $z \geq \max(x, y)$.

Alternative weighted attack principles for conjunction, formulated in the same manner, are:

$(\mathbf{Ł}^w_{\geq}.\wedge)$ If $F \overset{x}{\longrightarrow} A$, $F \overset{y}{\longrightarrow} B$, and $F \overset{z}{\longrightarrow} A \wedge B$, then $z \geq \min(1, x + y)$.

$(\mathbf{P}^w_{\geq}.\wedge)$ If $F \overset{x}{\longrightarrow} A$, $F \overset{y}{\longrightarrow} B$, and $F \overset{z}{\longrightarrow} A \wedge B$, then $z \geq x + y - xy$.

As the labels indicate, these principles are essential for obtaining an argumentation based semantics for Gödel logic G, Łukasiewicz logic Ł and Product logic P, respectively. Moreover the subscript '$\geq$' attached to these letters indicate that upper bounds for the weight of attacks of conjunctive claims (in terms of weights of attacks on conjuncts) are formulated here. In fact, also principles expressing matching lower bounds are needed to characterize the three mentioned t-norm based fuzzy logics. Correspondingly, we use $(\mathbf{G}^w_{\leq}.\wedge)$, $(\mathbf{Ł}^w_{\leq}.\wedge)$, and $(\mathbf{P}^w_{\leq}.\wedge)$ to refer to the principles that arise by just replacing '$\geq$' by '$\leq$' in the respective constraint.

As already indicated, in contrast to the qualitative case of Section 3, we do not have to involve threshold values in interpreting a weighted attack relation, but simply identify the weight with which F attacks A with the conditional probability that A does not hold, given that F holds. More formally, our probabilistic semantics interprets $F \overset{w}{\longrightarrow} A$ by $p(\neg A|F) = w$. (Remember that this is only viable if we exclude the possibility that F is a logical contradiction.) Accordingly, the above versions of weighted attack principles translate into the following statements.

$(\mathbf{G}^w_{\geq}.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \geq \max(x, y)$.

$(\mathbf{Ł}^w_{\geq}.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \geq \min(1, x + y)$.

$(\mathbf{P}^w_{\geq}.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \geq x + y - xy$.

$(\mathbf{G}^w_{\leq}.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \leq \max(x, y)$.

$(\mathbf{Ł}^w_{\leq}.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \leq \min(1, x + y)$.

$(\mathbf{P}^w_{\leq}.\wedge)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \wedge B)|F) \leq x + y - xy$.

According to our probability based interpretation we obtain the following classification of these principles.

Proposition 3. *The principles $(\mathbf{G}^w_{\geq}.\wedge)_p$ and $(\mathbf{Ł}^w_{\leq}.\wedge)_p$ hold in the sense of coherence-based probability logic. However, $(\mathbf{Ł}^w_{\geq}.\wedge)_p$, $(\mathbf{P}^w_{\geq}.\wedge)_p$, $(\mathbf{G}^w_{\leq}.\wedge)_p$, and $(\mathbf{P}^w_{\leq}.\wedge)_p$ do not hold for all coherent probability assessments.*

Proof. Remember that we assume that all involved propositions are classical. Therefore $\neg(A \wedge B)$ is equivalent to $\neg A \vee \neg B$, and hence the well known Fréchet inequalities (generalized to conditional probabilities) for logical disjunction yield $(\mathbf{G}^w_{\geq}.\wedge)_p$ and $(\mathbf{Ł}^w_{\leq}.\wedge)_p$.

The four other principles can all be violated:

$(\mathbf{Ł}^w_{\geq}.\wedge)_p$, $(\mathbf{P}^w_{\geq}.\wedge)_p$: Let $A = B$ and $p(\neg A|F) = p(\neg B|F) = 0.5$. Then $p(\neg(A \wedge B)|F) = p(\neg(A \wedge A)|F) = p(\neg A|F) = 0.5$, which is strictly smaller than $\min(1, 0.5 + 0.5) = 1$, but also strictly smaller than $0.5 + 0.5 - 0.5^2 = 0.75$.

$(\mathbf{G}^w_{\leq}.\wedge)_p$, $(\mathbf{P}^w_{\leq}.\wedge)_p$: Let $A = \neg B$ and $p(\neg A|F) = p(\neg B|F) = 0.5$. Then $p(\neg(A \wedge B)|F) = p(\neg(A \wedge \neg A)|F) = p(\neg\bot|F) = p(\top|F) = 1$, which is strictly larger than $\max(0.5, 0.5) = 0.5$ and strictly larger than $0.5 + 0.5 - 0.5^2 = 0.75$. □

Note that $(\mathbf{G}^w_{\geq}.\wedge)_p$ and $(\mathbf{Ł}^w_{\leq}.\wedge)_p$ define the best possible coherent lower and upper bounds, respectively. The principles $(\mathbf{Ł}^w_{\geq}.\wedge)_p$, $(\mathbf{P}^w_{\geq}.\wedge)_p$, $(\mathbf{G}^w_{\leq}.\wedge)_p$, and $(\mathbf{P}^w_{\leq}.\wedge)_p$, which do not hold under coherence, are not simply unjustifiable from a probabilistic point of view. They rather apply only to specific cases. The following corresponding propositions are straightforward.

Proposition 4. *Under the assumption that $p(A|F)$ and $p(B|F)$ are independent, $(\mathbf{P}^w_{\geq}.\wedge)_p$ and $(\mathbf{P}^w_{\leq}.\wedge)_p$ hold.*

Proposition 5. *Under the assumption that $A \models B$ or $B \models A$ $(\mathbf{G}^w_{\leq}.\wedge)_p$ holds.*

Proposition 6. *Under the assumption that $A \models \neg B$ or $B \models \neg A$ $(\mathbf{Ł}^w_{\leq}.\wedge)_p$ holds.*

The picture obtained for attack principles involving disjunction is, of course, dual to that just outlined for conjunction. The Fréchet inequalities justify the following two principles:

$(\mathbf{G}^w_{\leq}.\vee)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \vee B)|F) \leq \min(x, y)$.

$(\mathbf{Ł}^w_{\geq}.\vee)_p$ If $p(\neg A|F) = x$ and $p(\neg B|F) = y$ then $p(\neg(A \vee B)|F) \geq \max(0, x+y-1)$.

Other principles are justified according to the probabilistic semantics of argument attack only under additional assumptions about the (in)dependence of involved propositions.

For negation the probability semantics directly justifies the following attack principle, that combines and generalizes the qualitative principles **(A.¬)** and **(C.¬)**.

$(\mathbf{AC}^w.\neg)$ $F \xrightarrow{x} A$ if and only $F \xrightarrow{1-x} \neg A$.

Regarding implication, one may of course extract corresponding principles from the above mentioned ones, under the stipulation that $A \supset B$ is understood, classically, as equivalent to $\neg A \vee B$. But, once more, let us emphasize that it were actually more adequate to model (informal) implication as a conditional. This leads to the tricky and, as yet, only partially explored terrain of iterated conditional probabilities; thus providing a challenging topic for future research.

Draft names	Task/argument form	Task
(**A.**$\wedge$)	if $A \xrightarrow{x} B$, then $A \xrightarrow{[x,1]} (B \wedge C)$	B2,C4
(**C.**$\wedge$)	if $A \xrightarrow{x} (B \wedge C)$, then $A \xrightarrow{[0,x]} B$	A1,C1
(**A.**$\vee$)	if $A \xrightarrow{x} (B \vee C)$, then $A \xrightarrow{[x,1]} B$	A2,C3
(**C.**$\vee$)	if $A \xrightarrow{x} B$, then $A \xrightarrow{[0,x]} (B \vee C)$	B3,C6
Irrelevant premise	if $A \xrightarrow{x} B$ and $C \models B$ then $A \xrightarrow{x} B$	A3,C5
(**B.**$\neg$')	if $A \xrightarrow{x} B$, then $A \xrightarrow{1-x} \neg B$	B1,C2
Complement	if $A \xrightarrow{x} \neg B$, then $A \xrightarrow{1-x} B$	A4,C7
(**B.**$\neg$)	"if $A \longrightarrow B$, then $\neg(A \longrightarrow \neg B)$" is true	B11,C18
(**B.**$\neg$")	"if $A \longrightarrow \neg B$, then $\neg(A \longrightarrow B)$" is true	B12,C19
Narrow negation	if $A \xrightarrow{x} B$, then $A \xrightarrow{1-x} \neg B$	A7, B5,C11
(A.$\bot$**)**	$A \xrightarrow{1} (B \wedge \neg B)$	B4,C9
(A.$\top$**)**	$A \xrightarrow{0} (B \vee \neg B)$	B8,C15
Aristotle's thesis 1	$\neg(\neg A \longrightarrow A)$ is false	B6,C12
Aristotle's thesis 2	$\neg(A \longrightarrow \neg A)$ is false	A5,C8
Abelard's thesis	$\neg((A \longrightarrow B) \wedge (A \longrightarrow \neg B))$ is true	B7,C14
Reflexivity	$A \xrightarrow{0} A$	A6,C10
Contingent attack	$A \xrightarrow{[0,1]} B$	A8,C13
ProbToAttack	if $P(B\|A) = x$, then $A \xrightarrow{x} \neg B$	A10,B9,C17
AttackToProb	if $A \xrightarrow{x} B$, then $P(\neg B\|A) = x$	A9,B10
AttackToProb'	if $A \xrightarrow{x} B$, then $P(B\|A) = 1 - x$	C16
ProbToAttack'	if $P(B\|A) = x$, then $A \xrightarrow{1-x} B$	C17

Table 1: Task names/argument forms of the task sets with closed (i.e., conditions A and B) and open (i.e., C) response format. "$A \xrightarrow{x} B$" denotes "*A attacks by strength x the assertion B*", where x can be point- or interval-valued.

5 Experiment

In this section we explore the psychological plausibility of the proposed approach. Coherence-based probability logic received empirical support in recent years (e.g., [14, 16, 17, 18]). However, principles governing the strength of attacks have not yet been investigated empirically (neither within nor outside the coherence framework).

Participants The sample consists of 139 students of the Technical University of Vienna (18 females, 116 males, and 5 who chose not to reveal their gender) with a mean age of 21.1 years ($SD = 3.2$). Only German native speakers were included in the data analysis. Seven participants were excluded from the analysis because of missing data in the target tasks. Most students were in their second semester and did not receive a thorough training in logic yet.

On the average, the participants rated the overall task clearness and difficulty on an intermediate level ($M = 4.9$ and $M = 4.3$, respectively, on a rating scale out of 10). This reflects the fact that since our study aims to explore the interpretation of attack principles, the participants had first to reason towards how to interpret

the tasks and then, after fixing their interpretation, to draw conclusions based on their interpretation. This can also explain why the participants were not highly confident in the correctness of their solutions ($M = 4.1$ out of 10) even if in general they tend to like solving mathematical puzzles ($M = 7.5$ out of 10).

Method and materials Each participant was administered a DIN-A4 page, containing an introduction on the first page and the target tasks on both pages. There were three between-participant conditions, two with multiple-choice (A: $n_1 = 44$ and B: $n_2 = 48$) and one with an open choice response format (C: $n_3 = 47$). After showing how to express the degree of attack from a scale form 0 to 10 and that claims can also be compounded (like **[A and B]**), the participants were presented with those tasks which are described in Table 1. For example, Task A1 presents the antecedent of a conditional: "If **A** attacks with **exactly** the strength **7** the claim **B**, then ...". Then seven consequent candidates were presented, which completed the conditional. Eight consequents were of the form "...attacks **A** with [M] with the strength [S] the claim **B**", where "[M]" indicates a precise value ("exactly"), a lower ("at least"), or an upper bound ("at most") on the strength [S]. [S] was either 0, 3, 7, or 10. All possible point and interval options were formulated in ascending order (see Table 2 for the attack strength options we used). Except for the interval $[0, 10]$ we used "nothing follows about how strong ...attacks ...", as the ninth response option within each task. The participants were asked to tick for each of nine items whether the according sentence is correct ("*richtig*") or false ("*falsch*"). In the open response format condition C, the participants were instructed to fill in "exactly", "at least", or "at most", the value of the strength, and additionally had to mark the strength of attack (either as a point value or an interval) on a scale as introduced in the introduction. In all conditions, those tasks which were not formulated directly in terms of a conditionals, the instruction required to choose among "true", "false", or "undetermined" by ticking one corresponding box (e.g, A6, B4, B6, or B11 ; see Table 1).

The experiment took place during the last part of the first lecture on "formal modeling". The three conditions were administered in a systematically alternated way to reduce the chance of plagiarized responses.

Results and discussion The main results are presented in tables 2–6. First we observe that most people are unaware of the best possible coherent bounds (marked in **bold**). Responses which are within the optimal coherent bounds are of course also coherent, like in task A1 where 45% of the participants responded that "precisely 7" is correct. In this task, 43% responded that the interval "at most 7" is correct, which corresponds to the coherent interval. Second, we observe that compared to direct tests of coherence-based probability logic (e.g., [14, 16, 17, 18]), the agreement between the predictions concerning the quantitative attack principles and the participant's responses are modest, especially for the conditions with closed response formats (A and B). For the condition C, more than half of the participants responded by at least a coherent lower or a coherent upper bound as predicted (see

Task	0	[0,3]	3	[0,7]	[3,10]	7	[7,10]	10	nf
A1	0.00	0.00	0.00	**43.18**	18.18	45.45	18.18	0.00	31.82
A2	0.00	0.00	0.00	63.64	6.82	25.00	**9.09**	0.00	34.09
A3	0.00	2.27	0.00	25.00	18.18	**93.18**	27.27	0.00	4.55
A4	20.45	18.18	**18.18**	11.36	2.27	2.27	0.00	0.00	59.09
A7	15.91	22.73	**20.45**	13.64	6.82	9.09	0.00	0.00	52.27
A8	6.82	4.55	4.55	6.82	4.55	4.55	4.55	4.55	**88.64**
A9	2.27	13.64	22.73	2.27	9.09	**13.64**	6.82	4.55	56.82
A10	4.55	4.55	**13.64**	2.27	9.09	11.36	11.36	2.27	63.64

Table 2: Percentages of "correct" responses concerning the point valued/interval attack strength options in condition A ($n_1 = 44$). The response options of A9 were normalized to probability values. "nf" denotes "nothing follows". Best possible coherent response options are in **bold** (for predictions see Table 1).

Task	0	[0,3]	3	[0,7]	[3,10]	7	[7,10]	10	nf
B1	8.33	31.25	**29.17**	2.08	4.17	2.08	0.00	0.00	43.75
B2	2.08	4.17	2.08	22.92	18.75	16.67	**20.83**	0.00	39.58
B3	2.08	4.17	2.08	**27.08**	18.75	25.00	33.33	4.17	27.08
B5	8.33	31.25	**29.17**	0.00	4.17	0.00	2.08	4.17	45.83
B9	4.17	14.58	16.67	8.33	0.00	**2.08**	4.17	0.00	62.50
B10	2.08	12.50	14.58	8.33	4.17	**20.83**	4.17	0.00	47.92

Table 3: Percentages of "correct" responses in condition B ($n_2 = 48$). The response options of B10 were normalized to probability values. See also caption of Table 2.

median values in Table 5). Concerning the seven forced choice tasks in condition C, the most frequent responses were consistent with our coherence-based predictions in five tasks (see Table 6). In tasks C9 and C15 people chose incoherent responses, which involve contradictions and tautologies, which appear difficult to interpret in the context of principles about argument strength. We observed an analogous effect in the corresponding tasks B4 and B8 in the closed response format condition (see Table 4).

The Contingent attack tasks serve to check whether people read the tasks carefully. The Irrelevant premise task was intended to test (**A.gen**) but due to a systematic error in the translation of this argument form into the corresponding tasks, we use it now as a consistency check. In both tasks almost all participants responded as expected. The results of those tasks, which serve to explore directly the connection between probability and strength of attack (i.e., ProbToAttack, AttackToProb, and AttackToProb) were disappointing in the closed response format conditions A and B. In the open response format task C16, which investigates AttackToProb, the majority of participants responded as predicted. In task C17, which investigates ProbToAttack, the majority of only the lower bound responses were coherent. Again, participants scored better in the open response format condition compared to the closed one.

	A5	A6	B4	B6	B7	B8	B11	B12
false	**43.18**	**40.91**	31.25	**47.92**	41.67	**16.67**	31.25	31.25
correct	31.82	22.73	**25.00**	35.42	**31.25**	56.25	**39.58**	**35.42**
undetermined	25.00	36.36	43.75	16.67	27.08	27.08	29.17	33.33

Table 4: Percentages of responses in conditions A ($n_1 = 44$) and B ($n_2 = 48$). Best possible coherent response options are in **bold** (see Table 1).

	C1l	C1u	C2l	C2u	C3l	C3u	C4l	C4u	C5l	C5u	C6l
	0	**.70**	**.30**	**.30**	**.70**	**1**	**.70**	**1**	**.70**	**.70**	**0**
a	.42	.70	.16	.43	.25	.74	.37	.83	.63	.73	.31
b	.33	.20	.20	.36	.33	.18	.33	.22	.22	.08	.35
c	.70	.70	.00	.30	.00	.70	.30	1.00	.70	.70	.00
	C6u	C7l	C7u	C11l	C11u	C13l	C13u	C16l	C16u	C17l	C17u
	.70	**.30**	**.30**	**.30**	**.30**	**0**	**1**	**.30**	**.30**	**.30**	**.30**
a	.77	.17	.51	.14	.48	.24	.92	.31	.51	.28	.57
b	.18	.21	.40	.19	.38	.33	.21	.33	.34	.28	.34
c	.70	.00	.30	.00	.30	.00	1.00	.30	.30	.30	.70

Table 5: Mean (a), standard deviations (b), and medians (c) of lower (L) and upper (U) bound responses in condition C ($n_3 = 47$). Except for the probability responses to task C16, all values are normalized to the value range $[0, 1]$. Best possible coherent response options are in **bold** (see Table 1).

6 Concluding remarks

We showed how the coherence approach to probability can serve to guide the rational selection of qualitative and quantitative attack principles. More research is needed to deepen and to generalize our formal results: e.g., by interpreting implication by conditional probability (or by previsions in conditional random quantities) or by generalizations to fuzzy events. We also presented an experiment to explore the psychological plausibility of the proposed approach. While we are convinced that our approach is intuitive and plausible, we were surprised by the relatively heterogeneous results. Open response format tasks turned out the be more appropriate to investigate quantitative attack principles. The heterogeneous agreement between the predictions and the responses could be caused by various factors including (i) lower data quality in a lecture hall experiment compared to individual testing, (ii) different response formats, and (iii) possible confusions caused by the negations involved in the probabilistic semantics of the attack relations (i.e., $p(\neg B|A)$ should be high in order that $A \longrightarrow B$ holds). Future experimental work is needed to further explore the psychological plausibility of formal attack principles.

References

[1] P. Besnard and A. Hunter. *Elements of argumentation.* MIT Press Cambridge, 2008.

[2] G. Coletti and R. Scozzafava. *Probabilistic logic in a coherent setting.* Kluwer, 2002.

	C8	C9	C10	C14	C15	C18	C19
false	**51.06**	23.40	**76.60**	29.79	**14.89**	17.02	12.77
true	19.15	**27.66**	8.51	**42.55**	34.04	**48.94**	**53.19**
undetermined	29.79	48.94	14.89	27.66	51.06	34.04	34.04

Table 6: Percentages of responses to forced choice tasks in condition C ($n_3 = 47$). Best possible coherent response options are in **bold** (see Table 1).

[3] E.A. Corsi and C.G. Fermüller. Logical argumentation principles, sequents, and nondeterministic matrices. In A. Baltag, J. Seligman, and T. Yamada, editors, *LORI 2017*, volume 10455 of *LNCS*, pages 422–437, Berlin, 2017. Springer.

[4] E.A. Corsi and C.G. Fermüller. Connecting fuzzy logic and argumentation frames via logical attack principles, submitted.

[5] Sylvie Coste-Marquis, Sébastien Konieczny, Pierre Marquis, and Mohand Akli Ouali. Weighted attacks in argumentation frameworks. In *13th International Conference on the Principles of Knowledge Representation and Reasoning (KR'12)*. AAAI Press, 2012.

[6] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intelligence*, 77(9):321–357, 1995.

[7] J Michael Dunn and Greg Restall. Relevance logic. In *Handbook of Philosophical Logic*, volume 6, pages 1–128. Springer, 2002.

[8] Paul E Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486, 2011.

[9] A. Gilio. Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence*, 34:5–34, 2002.

[10] A. Gilio, N. Pfeifer, and G. Sanfilippo. Transitivity in coherence-based probability logic. *Journal of Applied Logic*, 14:46–64, 2016.

[11] U. Hahn and M. Oaksford. The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, 114(3):704–732, 2007.

[12] Ingebrigt Johansson. Der Minimalkalkül, ein reduzierter intuitionistischer Formalismus. *Compositio Mathematica*, 4:119–136, 1937.

[13] M. Oaksford, N. Chater, and U. Hahn. Human reasoning and argumentation: The probabilistic approach. In J. Adler and L. Rips, editors, *Reasoning: Studies of human inference and its foundations*. Cambridge University Press, Cambridge, in press .

[14] N. Pfeifer. The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning*, 19(3–4):329–345, 2013.

[15] N. Pfeifer. Reasoning about uncertain conditionals. *Studia Logica*, 102(4):849–866, 2014.

[16] N. Pfeifer and G. D. Kleiter. Coherence and nonmonotonicity in human reasoning. *Synthese*, 146(1-2):93–109, 2005.

[17] N. Pfeifer and G. D. Kleiter. Framing human inference by coherence based probability logic. *Journal of Applied Logic*, 7(2):206–217, 2009.

[18] N. Pfeifer and L. Tulkki. Conditionals, counterfactuals, and rational reasoning. An experimental study on basic principles. *Minds and Machines*, 27(1):119–165, 2017.

[19] G. Sanfilippo, N. Pfeifer, and A. Gilio. Generalized probabilistic modus ponens. In A. et al. Antonucci, editor, *ECSQUARU 2017*, volume 10369 of *LNCS*, pages 480–490. Springer, 2017.

[20] G. Sanfilippo, N. Pfeifer, D. E. Over, and A. Gilio. Probabilistic inferences from conjoined to iterated conditionals. *International Journal of Approximate Reasoning*, 93:103–118, 2018.

[21] F. Zenker, editor. *Bayesian argumentation: The practical side of probability*. Synthese Library (Springer), Dordrecht, 2013.

Gradient Descent Parameter Learning of Bayesian Networks under Monotonicity Restrictions

Martin Plajner[1,2]
plajner@utia.cas.cz
Faculty of Nuclear Sciences and Physical Engineering[1],
Czech Technical University, Prague
Trojanova 13, Prague,
120 00, Czech Republic

Jiří Vomlel[2]
Institute of Information Theory and Automation[2],
Czech Academy of Sciences,
Pod vodárenskou věží 4, Prague 8, 182 08, Czech Republic

Abstract

Learning parameters of a probabilistic model is a necessary step in most machine learning modeling tasks. When the model is complex and data volume is small the learning process may fail to provide good results. In this paper we present a method to improve learning results for small data sets by using additional information about the modelled system. This additional information is represented by monotonicity conditions which are restrictions on parameters of the model. Monotonicity simplifies the learning process and also these conditions are often required by the user of the system to hold.

In this paper we present a generalization of the previously used algorithm for parameter learning of Bayesian Networks under monotonicity conditions. This generalization allows both parents and children in the network to have multiple states. The algorithm is described in detail as well as monotonicity conditions are.

The presented algorithm is tested on two different data sets. Models are trained on differently sized data subsamples with the proposed method and the general EM algorithm. Learned models are then compared by their ability to fit data. We present empirical results showing the benefit of monotonicity conditions. The difference is especially significant when working with small data samples. The proposed method outperforms the EM algorithm for small sets and provides comparable results for larger sets.

1 Introduction

In our research we address Computerized Adaptive Testing (CAT) [1, 13]. CAT is a concept of testing latent student abilities which allows us to create shorter tests, asking less questions in a shorter time while keeping the same level of information. This task is performed by asking the right questions for each individual student. Questions are selected based on a student model. In common practice experts often use Item Response Theory models [10] (IRT) which are well explored and have been in use for a long time. Nevertheless, we have focused our attention on a different family of models to model a student using Bayesian Networks (BNs) since they offer more options in the modelling process. It is for example possible to model more complex influences between skills and questions as BNs are not limited to connecting each skill with each question as well as we can introduce connections between skills themselves.

During our research we noticed that there are certain conditions which should be satisfied in this specific modelling task. We especially focused on monotonicity conditions. Monotonicity conditions incorporate qualitative influences into a model. These influences restrict conditional probabilities inside the model in a specific way to avoid unwanted behavior. Monotonicity in Bayesian Networks has been discussed in the literature for a long time. It is addressed, selecting the most relevant to our topic, by [14, 3] and more recently by ,e.g., [11, 5]. Monotonicity restrictions are often motivated by reasonable demands from model users. In our case of CAT it means we want to guarantee that students having certain skills will have a higher probability of answering questions correctly.

Certain types of models include monotonicity naturally by the way they are constructed. In the case of general BNs this is not true. In order to satisfy these conditions we have to introduce restrictions to conditional probabilities during the process of parameter learning.

In our previous work we first showed that monotonicity conditions are uself in the context of CAT [8]. Later we applied these conditions to Bayesian Network [9]. In this article we extend our earlier presented gradient descent optimum search method for BN parameter learning under monotonicity conditions. The last article covers only specific BNs. It works solely with binary children variables in the model (yes/no answers in terms of CAT). The extension we present in this article provides a tool to include monotonicity in BN models with multiple-state children nodes. Additionally, in this article we perform experiments on a new dataset. It is consists of data from the Czech high school state final exam. This data source contains a large volume of reliable data, and it is very useful for the empirical verification of our ideas.

We implemented the new method in R language and performed experimental verification of our assumptions. We used two data sets. The first one, a synthetic data set, is generated from artificial models satisfying monotonicity conditions. The second one, an empirical data set, is formed by data from the Czech high school final exam. Experiments were performed on these data sets also with the

ordinary EM learning without monotonicity restrictions in order to compare these two approaches.

The structure of this article is as follows. First, we establish our notation and describe monotonicity conditions in detail in Section 2. Next, we present the extended method in Section 3. In Section 4 of this paper, we take a closer look at the experimental setup and present results of our experiments. The last section contains an overview and a discussion of the obtained results.

2 BN Models and Monotonicity

2.1 Notation

In this article we use the new gradient descent method for BNs which are used to model students in the domain of CAT. Details about BNs can be found, for example, in [7, 6]. We restrict ourselves to the BNs that have two levels. In compliance with our previous articles, variables in the parent level are addressed as skill variables S. The children level contains questions variables X. Examples of network structures, which we also used for experiments, are shown in Figures 1 and 2.

- We use the symbol $\boldsymbol{X}$ to denote the multivariable $(X_1, \ldots, X_n)$ taking states $\boldsymbol{x} = (x_1, \ldots, x_n)$. The total number of question variables is n, the set of all indexes of question variables is $\boldsymbol{N} = \{1, \ldots, n\}$. Question variables' individual states are $x_{i,t}, t \in \{0, \ldots, n_i\}$ and they are observable. Each question can have a different number of states, the maximum number of states over all variables is $N^{max} = \max_i(n_i) + 1$. States are integers with natural ordering specifying the number of points obtained in the $i-th$ question[1].
- We use the symbol $\boldsymbol{S}$ to denote the multivariable $(S_1, \ldots, S_m)$ taking states $\boldsymbol{s} = (s_1, \ldots, s_m)$. The set of all indexes of skill variables is $\boldsymbol{M} = \{1, \ldots, m\}$. Skill variables have a variable number of states, the total number of states of a variable S_j is m_j, and individual states are $s_{j,k}, k \in \{1, \ldots, m_j\}$. The variable $\boldsymbol{S}^i = \boldsymbol{S}^{pa(i)}$ stands for a multivariable containing only parent variables of the question X_i. Indexes of these variables are $\boldsymbol{M}^i \subseteq \boldsymbol{M}$. The set of all possible state configurations of $\boldsymbol{S}^i$ is $Val(\boldsymbol{S}^i)$. Skill variables are unobservable.

The BN has CPT parameters for all questions $X_i, i \in \boldsymbol{N}, \boldsymbol{s}^i \in Val(\boldsymbol{S}^i)$ which define conditional probabilities as

$$P(X_i = t | \boldsymbol{S} = \boldsymbol{s}) = \theta^t_{i,\boldsymbol{s}^i} \ ,$$

and for all parent variables $S_j, j \in \boldsymbol{M}$ as

$$P(S_j = s_j) = \tilde{\theta}_{j,s_j} \ .$$

[1]The interpretation of points is very complex and has to be viewed as per question because we use the CAT framework. In this context getting one point in one question is not the same as one point in another.

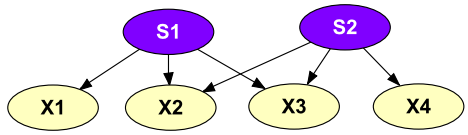


Figure 1: An artificial BN model

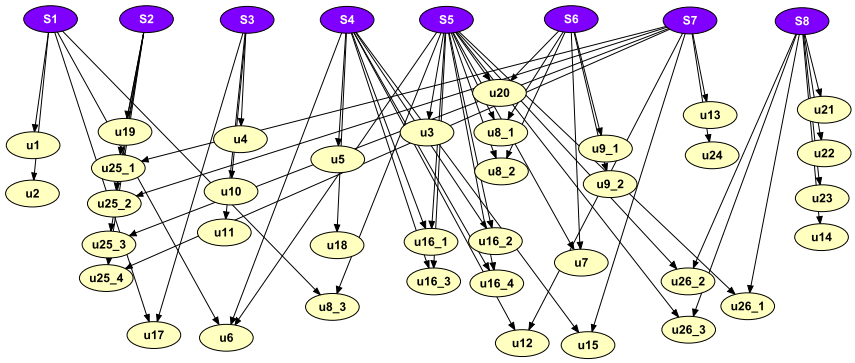


Figure 2: A BN model for CAT

From the definition above it follows that parameters are constrained to be between zero and one and to sum up to one. For question variable the condition is $\sum_{t=0}^{n_i} \theta^t_{i,\boldsymbol{s}^i} = 1,\ \forall i, \boldsymbol{s}^i$ and for parent variables it is $\sum_{s_j} \tilde{\theta}_{j,s_j} = 1,\ \forall j$. To remove this condition for the later use in the gradient method we reparametrize parameters

$$\begin{aligned} \theta^t_{i,\boldsymbol{s}^i} &= \frac{exp(\mu^t_{i,s_i})}{\sum_{t'=0}^{n_i} exp(\mu^{t'}_{i,s_i})} \\ \tilde{\theta}_{j,s_j} &= \frac{exp(\tilde{\mu}_{j,s_j})}{\sum_{s'_j=1}^{m_i} exp(\tilde{\mu}_{j,s'_j})} \quad . \end{aligned}$$

The set of all question parameters $\theta^t_{i,\boldsymbol{s}^i}$ and all skills parameters $\tilde{\theta}_{j,s_j}$ is $\boldsymbol{\theta}$ without the reparametrization and $\boldsymbol{\mu}$ with the reparametrization.

2.2 Monotonicity

The concept of monotonicity in BNs has been discussed in the literature since the last decade of the previous millennium [14, 3]. Later its benefits for BN parameter learning were addressed, for example, by [12, 2]. This topic is still active, e.g., [4, 11, 5].

We consider only variables with states from $\mathbb{N}_0$ with their natural ordering, i.e.,

the ordering of states of skill variable S_j for $j \in \boldsymbol{M}$ is

$$s_{j,1} \prec \ldots \prec s_{j,m_j} \ .$$

A variable S_j has a monotone effect on its child X_i if for all $k, l \in \{1, \ldots, m_j\}, t' \in \{0, \cdots, n_i\}$:

$$s_{j,k} \preceq s_{j,l} \quad \Rightarrow \quad \sum_{t=0}^{t'} P(X_i = t | S_j = s_{j,k}, \boldsymbol{s}) \ \geq \ \sum_{t=0}^{t'} P(X_i = t | S_j = s_{j,l}, \boldsymbol{s})$$

and antitone effect:

$$s_{j,k} \preceq s_{j,l} \quad \Rightarrow \quad \sum_{t=0}^{t'} P(X_i = t | S_j = s_{j,k}, \boldsymbol{s}) \ \leq \ \sum_{t=0}^{t'} P(X_i = t | S_j = s_{j,l}, \boldsymbol{s}) \ ,$$

where $\boldsymbol{s}$ is a configuration of remaining parents of question i without S_j. For each question $X_i, i \in \boldsymbol{M}$ we denote by $\boldsymbol{S}^{i,+}$ the set of parents with a monotone effect and by $\boldsymbol{S}^{i,-}$ the set of parents with an antitone effect.

The conditions above are defined for states of question variable X_i in the set $\{0, \cdots, (n_i - 1)\}$. Given the property of conditional probabilities, i.e.

$$\theta^{n_i}_{i,\boldsymbol{s}^i} = 1 - \sum_{t=0}^{n_i - 1} \theta^{t}_{i,\boldsymbol{s}^i} \ ,$$

it holds for the state n_i in the form for monotonic:

$$s_{j,k} \preceq s_{j,l} \quad \Rightarrow \quad P(X_i = n_i | S_j = s_{j,k}, \boldsymbol{s}) \ \leq \ P(X_i = n_i | S_j = s_{j,l}, \boldsymbol{s})$$

and for antitonic:

$$s_{j,k} \preceq s_{j,l} \quad \Rightarrow \quad P(X_i = n_i | S_j = s_{j,k}, \boldsymbol{s}) \ \geq \ P(X_i = n_i | S_j = s_{j,l}, \boldsymbol{s})$$

Next, we create a partial ordering $\preceq_i$ on all state configurations of parents $\boldsymbol{S}^i$ of the i-th question, where for all $\boldsymbol{s}^i, \boldsymbol{r}^i \in Val(\boldsymbol{S^i})$:

$$\boldsymbol{s}^i \quad \preceq_i \quad \boldsymbol{r}^i \Leftrightarrow \left(s^i_j \preceq r^i_j, \ j \in \boldsymbol{S}^{i,+}\right) \text{ and } \left(r^i_j \preceq s^i_j, \ j \in \boldsymbol{S}^{i,-}\right) \ .$$

The monotonicity condition then requires that the probability of an incorrect answer is higher for a lower order parent configuration (chances of correct better answers increasing for higher ordered parents' states), i.e., for all $\boldsymbol{s}^i, \boldsymbol{r}^i \in Val(\boldsymbol{S^i}), k \in \{0, \ldots, (n_i - 1)\}$:

$$\boldsymbol{s}^i \preceq_i \boldsymbol{r}^i \quad \Rightarrow \quad \sum_{t=0}^{k} P(X_i = t | \boldsymbol{S}^i = \boldsymbol{s}^i) \ \geq \ \sum_{t=0}^{k} P(X_i = t | \boldsymbol{S}^i = \boldsymbol{r}^i) \ .$$

In our experimental part we consider only the monotone effect of parents on their children. The difference with antitone effects is only in the partial ordering.

3 Parameter Gradient Search with Monotonicity

To learn the parameter vector $\boldsymbol{\mu}$ we have developed a method based on gradient descent optimization. We follow the work of [2] where authors use a gradient descent method with exterior penalties to learn parameters. The main difference is that we consider models with hidden variables. In this article we generalize the method from [9] to multistate question variables.

We denote by $\boldsymbol{D}$ the set of indexes of question vectors. One vector $x^k, k \in \boldsymbol{D}$ corresponds to one student and an observation of i-th variable X_i is x_i^k. The number of occurrences of the k-th configuration vector in the data sample is d_k.

We use the model as described in Section 2 having unobserved parent variables and observed children variables. With sets $\boldsymbol{I}_t^k, t \in \{0, \ldots, N^{max}\}$ of indexes of questions answered with the point gain of t points, we define the following products based on observations in the k-th vector:

$$p^t(\boldsymbol{\mu}, \boldsymbol{s}, k) = \prod_{i \in \boldsymbol{I}_t^k} \frac{exp(\mu_{i,\boldsymbol{s}}^t)}{\sum_{t'=0}^{n_i} exp(\mu_{i,\boldsymbol{s}}^{t'})},\ t \in \{0, \cdots, N^{max}\}; \quad p_\mu(\boldsymbol{\mu}, \boldsymbol{s}) \quad = \quad \prod_{j=1}^{m} exp(\tilde{\mu}_{j,s_j}).$$

We work with the log likelihood of data modelled by BN with the parameter vector $\boldsymbol{\mu}$:

$$\begin{aligned} LL(\boldsymbol{\mu}) &= \sum_{k \in \boldsymbol{D}} d_k \cdot log \left(\sum_{\boldsymbol{s} \in Val(\boldsymbol{S})} \prod_{j=1}^{m} \frac{exp(\tilde{\mu}_{j,s_j})}{\sum_{s'_j=1}^{m_j} exp(\tilde{\mu}_{j,s'_j})} \cdot \prod_{t=0}^{N^{max}} p^t(\boldsymbol{\mu}, \boldsymbol{s}, k) \right) \\ &= \sum_{k \in \boldsymbol{D}} d_k \cdot log\Big(\sum_{\boldsymbol{s} \in Val(\boldsymbol{S})} p_\mu(\boldsymbol{\mu}, \boldsymbol{s}) \prod_{t=0}^{N^{max}} p^t(\boldsymbol{\mu}, \boldsymbol{s}, k) \Big) - N \cdot \sum_{j=1}^{m} log \sum_{s'_j=1}^{m_j} exp(\tilde{\mu}_{j,s'_j}) \ . \end{aligned}$$

In the gradient descent optimization we need partial derivatives to establish the gradient. The partial derivatives of $LL(\boldsymbol{\mu})$ with respect to $\mu_{i,\boldsymbol{s}^i}$ for $i \in \boldsymbol{N}, \boldsymbol{s}^i \in Val(\boldsymbol{S}^i)$ are

$$\frac{\delta LL(\boldsymbol{\mu})}{\delta \mu_{i,\boldsymbol{s}^i}^t} =$$

$$\sum_{k \in \boldsymbol{D}} d_k \cdot \quad \frac{I(t, i, \boldsymbol{s^i}, k) - (\sum_{t'=0}^{n_i} exp(\mu_{i,\boldsymbol{s}^i}^{t'}) - exp(\mu_{i,\boldsymbol{s}^i}^t)) \cdot p_\mu(\boldsymbol{\mu}, \boldsymbol{s}^i) \prod_{t=0}^{N^{max}} p^t(\boldsymbol{\mu}, \boldsymbol{s}, k)}{\sum_{t'=0}^{n_i} exp(\mu_{i,\boldsymbol{s}^i}^{t'}) \cdot \sum_{\boldsymbol{s} \in Val(\boldsymbol{S})} \Big(p_\mu(\boldsymbol{\mu}, \boldsymbol{s}) \prod_{t=0}^{N^{max}} p^t(\boldsymbol{\mu}, \boldsymbol{s}, k) \Big)},$$

$$\text{where} \qquad I(t, i, \boldsymbol{s^i}, k) = \begin{cases} exp(\mu_{i,\boldsymbol{s}^i}^t), & \text{if } t = k \\ 0, & \text{otherwise} \end{cases}$$

and with respect to $\tilde{\mu}_{i,l}$ for $i \in \boldsymbol{M}, l \in \{1, \ldots, m_i\}$ are

$$\frac{\delta LL(\boldsymbol{\mu})}{\delta \tilde{\mu}_{i,l}} = \sum_{k \in \boldsymbol{D}} d_k \cdot \frac{\sum_{\boldsymbol{s} \in Val(\boldsymbol{S})}^{s_i=l} p_\mu(\boldsymbol{\mu}, \boldsymbol{s}) \prod_{t=0}^{N^{max}} p^t(\boldsymbol{\mu}, \boldsymbol{s}, k)}{\sum_{\boldsymbol{s} \in Val(\boldsymbol{S})} p_\mu(\boldsymbol{\mu}, \boldsymbol{s}) \prod_{t=0}^{N^{max}} p^t(\boldsymbol{\mu}, \boldsymbol{s}, k)} - N \cdot \frac{exp(\tilde{\mu}_{i,l})}{\sum_{l'=1}^{m_i} exp(\tilde{\mu}_{k,l'})} .$$

3.1 Monotonicity Restriction

To ensure monotonicity we use a penalty function which penalizes solutions that do not satisfy monotonicity conditions

$$C(\theta_{i,\boldsymbol{s}^i}, \theta_{i,\boldsymbol{r}^i}, t', c) = exp(c \cdot (\sum_{t=0}^{t'} \theta^t_{i,\boldsymbol{r}^i} - \sum_{t=0}^{t'} \theta^t_{i,\boldsymbol{s}^i}))$$

for the log likelihood:

$$LL'(\boldsymbol{\theta}, c) = LL(\boldsymbol{\theta}) - \sum_{i \in \boldsymbol{N}} \sum_{\boldsymbol{s}^i \preceq_i \boldsymbol{r}^i} \sum_{t'=0}^{N^{max}} C(\theta_{i,\boldsymbol{s}^i}, \theta_{i,\boldsymbol{r}^i}, t', c),$$

and in the case of reparametrized parameters:

$$LL'(\boldsymbol{\mu}, c) = LL(\boldsymbol{\mu}) - \sum_{i \in \boldsymbol{N}} \sum_{\boldsymbol{s}^i \preceq_i \boldsymbol{r}^i} \sum_{t'=0}^{N^{max}} C(\frac{exp(\mu^t_{i,\boldsymbol{s}^i})}{\sum_{t'=0}^{n_i} exp(\mu^{t'}_{i,\boldsymbol{s}^i})}, \frac{exp(\mu^t_{i,\boldsymbol{r}^i})}{\sum_{t'=0}^{n_i} exp(\mu^{t'}_{i,\boldsymbol{r}^i})}, t', c),$$

where c is a constant determining the slope of the penalization function. The higher the value the more strict the penalization is. Theoretically, this condition does not ensure monotonicity but, practically, selecting high values of c results in monotonic estimates. If the monotonicity is not violated then the penalty value is close to zero. Otherwise, the penalty is raising exponentially fast. In our experiments we have used the value of $c = 200$ but any value higher than 100 provided almost identical results.

After adding the penalized part to the log likelihood, partial derivatives with respect to $\mu_{i,l}$ remain unchanged. Partial derivatives with respect to $\mu^t_{i,\boldsymbol{s}^i}$ change. The reparametrization causes the derivatives to become very complex. Due to limited space in this paper we do not include their full description here.

Using the penalized log likelihood, $LL'(\boldsymbol{\mu}, c)$, and its gradient $\nabla(LL'(\boldsymbol{\mu}, c))$ we can use standard gradient descent optimization methods to find the paramters of BN models.

4 Experiments

We designed tests to verify our assumptions. We want to show that if we learn parameters of BNs with little amount data it is beneficial to use monotonicity

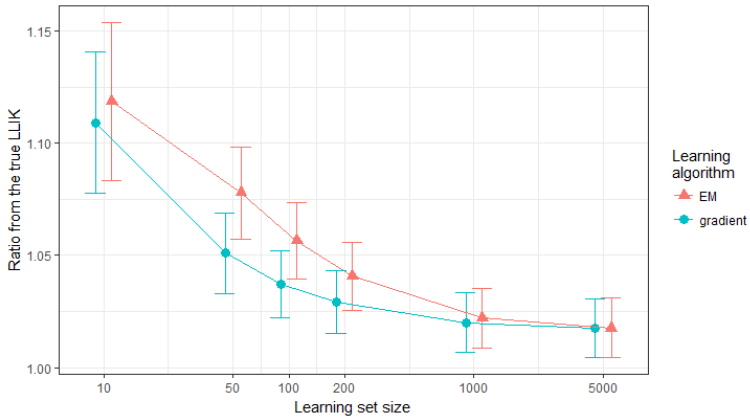


Figure 3: Artificial model: The ratio between the fitted and the real log likelihood (measured on the whole data set) obtained by models trained with EM and the restricted gradient methods for different training set sizes. Notice the logarithmic scale of the x axis. Curves are slightly misaligned in the direction of the x-axis to avoid overlapping.

constraints. We designed two experiments to test the method described above. The first one works with artificial (synthetic data); the other uses a real world empiric data sample.

Parameters are learned with our gradient method and the standard unrestricted EM algorithm. In both cases, we learn model parameters from subsets of data of different sizes. The quality of the parameter fit is measured by the log likelihood. The log likelihood is measured on the whole data set to provide results comparable between subsets of different sizes.

4.1 Artificial Model

The structure of the first model is shown in Figure 1. This model reflects the usual model structure used in CAT where there are two levels of variables, one level of questions, and one level of parents (skills). Parents S_1 and S_2 have 3 possible states and children X_1, X_2, X_3, X_4 also have three states. The model was set up with 10 different sets of parameters $\boldsymbol{\theta}_a^*$ satisfying the monotonicity conditions. Furthermore, every model produced 10 000 test cases.

To learn parameters of these models we drew random subsets of size d of 10, 50, 100, 200, 1 000, 5 000. Ten different sets for each size (indexed by b). Next, we created 10 initial starting points (indexed by c) for the model learning phase. The structure of both generating and learning models is the same and is shown in Figure 1. Starting parameter vectors $\boldsymbol{\theta}_b$ are randomized so that they satisfy the monotonicity conditions. Parameters of all parent variables are uniform. Starting

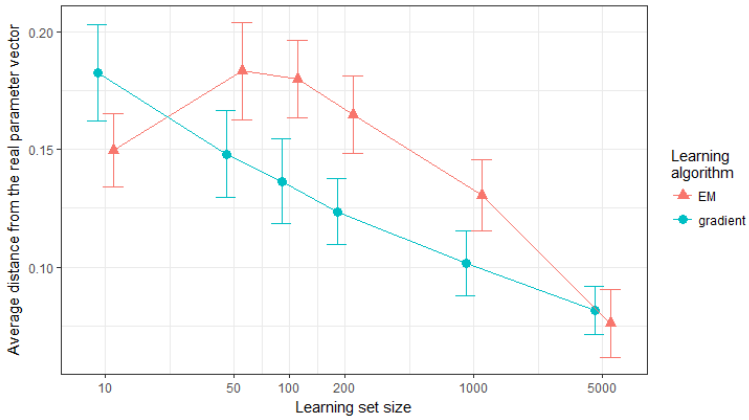


Figure 4: Artificial model: Mean parameter distance between real and fitted parameters in models trained with the EM and restricted gradient methods for different training set sizes. Notice the logarithmic scale of the x axis. Curves are slightly misaligned in the direction of the x-axis to avoid overlapping.

points are the same for both the EM and the gradient method alike. In this setup we have 10 different original models, 10 different observation subsets, and 10 different starting parameters, which gives 1 000 combinations for each set size. Each combination has a set of parameters $\boldsymbol{\theta}^d_{a,b,c}, a, b, c \in \{1, \ldots, 10\}$. We performed tests for all these combinations and the results are evaluated as follows.

We measure the log likelihood on the whole data set in order to keep results comparable. The resulting log likelihood after learning is compared with the log likelihood obtained with the real model and then averaged over all instances. This process gives us the average percentual difference between the original and fitted model. For the set size d:

$$LR^d = \frac{\sum_{a,b,c} \frac{LL(\boldsymbol{\theta}^*_a)}{LL(\boldsymbol{\theta}^d_{a,b,c})}}{1000}$$

Resulting valus for all set sizes are shown in Figure 3. In this artificial setup we are also able to measure the distance of learned parameters from the generating parameters. First we calculate an average error for each learned model:

$$e^d_{i,j} = \frac{|\boldsymbol{\theta}^*_a - \boldsymbol{\theta}^d_{a,b,c}|}{|\boldsymbol{\theta}|} \quad ,$$

where $||$ is the L1 norm. Next we average over all results in one set size d:

$$e^d = \frac{\sum_{a,b,c} e_{i,j}}{1000} \quad .$$

The summary of results is shown in Figure 4.

4.2 CAT Model

The second model is the model presented in Figure 2 and we use it for our CAT research. Parent variables $S_1, \ldots, S_8$ have 3 states and each one of them represents a particular student skill. Children nodes U_i are variables representing questions which have a various number of states (based on the evaluation of the specific question). This model was learned from data contained in the data sample collected from the Czech high school final exam[2]. The data set contains answers from over 20 000 students who took the test in the year 2015. We created the model structure based on our expert analysis and assigned skills to questions. To learn parameters we use random subsets of size of 10, 50, 100, and 500 cases of the whole sample. We drew 10 random sets for each size. Models were initiated with 10 different initial random starting parameters $\boldsymbol{\theta}_i$.

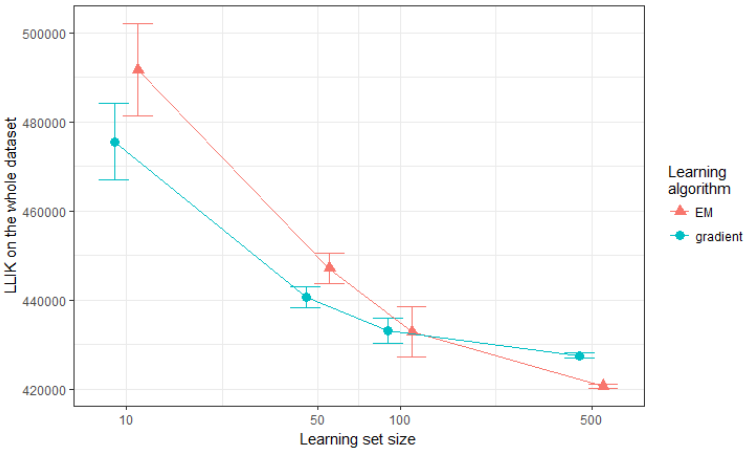


Figure 5: BN model for CAT empirical data: LLIK scored on the whole dataset for models trained with the EM and restricted gradient methods for different training set sizes. Notice the logarithmic scale of the x axis. Curves are slightly misaligned in the direction of the x-axis to avoid overlapping.

For the learned models we computed the log likelihood for the whole data set. These values are then averaged over all results of the same size $LL_A(k)$ similarly to the artificial model. Results are presented in Figure 5. In this case we cannot compare learned parameters because the real parameters are unknown.

[2]The test is accessible here (Czech language):`http://www.statnimaturita-matika.cz/wp-content/uploads/matematika-test-zadani-maturita-2015-jaro.pdf`

5 Conclusions

In this article we presented a new gradient based method for learning parameters of Bayesian Networks under monotonicity restrictions. The method was described and then tested on two data sets. In Figures 3 and 5 it is clearly visible that the newly proposed method provide better results than the general EM algorithm for small set sizes. When the size of learning set grows both method are getting more accurate and fitting data better. As we can see in results of the artificial model, both methods converge to the same point which is almost identical to the log likelihood of the model with real parameters. The speed of convergence is slower for the gradient method, nevertheless in the artificial case, it is not outperformed by the EM algorithm. In the case of empirical data, we can observe the same notion where for small set sizes the new gradient method is scoring better results. In this case EM is getting better log likelihood for larger data sets. This is caused by the fact that for these larger sets monotonicity restrictions start to make the learning process harder. For smaller sets they are showing the right path and guiding the learning process to a better solution. For larger sets they are restricting parameters and making the process harder. On the other hand, in case when we use the gradient method, we are working with learned model satisfying monotonicity conditions which may be desirable given its purpose.

This article shows that it is possible to benefit from monotonicity conditions. It presents the method to be used to learn parameter of BNs under these conditions. A possible extension of our work is to design a method which would use gradient descent optimization in a polytope defined by monotonicity conditions instead of using a penalty function. This approach has certain benefits as it ensures ending with strictly monotonic solution, on the other hand the current method allows small deviations from monotonicity if data strongly contradicts it.

Acknowledgements

This work was supported by the Czech Science Foundation (project No. 16-12010S) and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/198/OHK4/3T/14.

References

[1] R. G. Almond and R. J. Mislevy. Graphical Models and Computerized Adaptive Testing. *Applied Psychological Measurement*, 23(3):223–237, 1999.

[2] E. E. Altendorf, A. C. Restificar, and T. G. Dietterich. Learning from Sparse Data by Exploiting Monotonicity Constraints. *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI2005)*, 2005.

[3] J. Druzdzel and M. Henrion. Efficient Reasoning in Qualitative Probabilistic Networks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 548–553. AAAI Press, 1993.

[4] A. J. Feelders and L. van der Gaag. Learning Bayesian Network Parameters with Prior Knowledge about Context-Specific Qualitative Influences. *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI2005)*, 2005.

[5] A. R. Masegosa, A. J. Feelders, and L. van der Gaag. Learning from incomplete data in Bayesian networks with qualitative influences. *International Journal of Approximate Reasoning*, 69:18–34, 2016.

[6] T. D. Nielsen and F. V. Jensen. *Bayesian Networks and Decision Graphs (Information Science and Statistics)*. Springer, 2007.

[7] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., dec 1988.

[8] M. Plajner and J. Vomlel. Student Skill Models in Adaptive Testing. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 403–414. JMLR.org, 2016.

[9] M. Plajner and J. Vomlel. Monotonicity in Bayesian Networks for Computerized Adaptive Testing. In A. Antonucci, L. Cholvy, and O. Papini, editors, *ECSQARU 2017*, pages 125–134, Cham, 2017. Springer International Publishing.

[10] G. Rasch. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Danmarks Paedagogiske Institut, 1960.

[11] A. C. Restificar and T. G. Dietterich. Exploiting monotonicity via logistic regression in Bayesian network learning. Technical report, Corvallis, OR : Oregon State University, 2013.

[12] L. van der Gaag, H. L. Bodlaender, and A. J. Feelders. Monotonicity in Bayesian networks. *20th Conference on Uncertainty in Artificial Intelligence (UAI '04)*, pages 569–576, 2004.

[13] W. J. van der Linden and C. A. W. Glas. *Computerized Adaptive Testing: Theory and Practice*, volume 13. Kluwer Academic Publishers, 2000.

[14] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3):257–303, 1990.

An Expectation Operator for Belief Functions in the Dempster-Shafer Theory

Prakash P. Shenoy
School of Business
University of Kansas
pshenoy@ku.edu

Abstract

The main contribution of this paper is a new definition of an expectation operator for belief functions in the Dempster-Shafer (D-S) theory of evidence. Our definition shares many of the properties of the expectation operator in probability theory. Also, for Bayesian belief functions, our definition provides the same expected value as the probabilistic expectation operator. A traditional method of computing expected values of real-valued functions is to first transform a D-S belief function to a corresponding probability mass function, and then use the expectation operator for probability mass functions. Our expectation operator works directly with D-S belief functions. In general, our definition provides different expected values than, e.g., if we use probabilistic expectation using the pignistic transform or the plausibility transform of a belief function.

1 Introduction

The main goal of this paper is to propose an expectation operator for belief functions in the D-S theory of evidence [2, 6].

In probability theory, for discrete real-valued random variables characterized by a probability mass function (PMF), the expected value of X can be regarded as a weighted average of the states of X where the weights are the probabilities associated with the values. Our definition is similar. As we have probabilities associated with subsets of states, first we define the value of a subset as the weighted average of the states of the subset where the weights are the commonality values of the singleton states. Then the expected value of X is defined to be the weighted average of the values of the subsets where the weights are the commonality values of the subsets.

A traditional method of computing expectation of real-valued functions is to first transform a D-S belief function to a corresponding PMF, and then use the expectation operator for PMFs. Our expectation operator works directly with D-S belief functions. In general, our definition provides different expected values than, e.g., if we use the pignistic transform or the plausibility transform.

An outline of the remainder of the paper is as follows. In Section 2, we review the definition of expected value of a discrete real-valued random variable characterized by a PMF. Also, we review some of the main properties of the definition. In Section 3, we review the representations and operations of the D-S theory of belief functions. In Section 4, we provide our definition of the expected value of a real-valued random variable characterized by a commonality function. For a symbolic-valued random variable X, assuming we have a real-valued function g from the set of all non-empty subsets of the states of X, we also provide a definition of the expected value of g. Also, we show that our definition of expected value shares many of the properties of the probabilistic expected value, and we compare our definition with the probabilistic expectation using pignistic and plausibility transforms. Finally, in Section 5, we summarize and conclude.

2 Expected Values of Discrete Probability Distributions

In this section we briefly review the expectation operator for discrete random variables with finite state space whose behavior is described by probability mass functions.

2.1 Definition of probabilistic expectation

Suppose X is a discrete real-valued random variable with a finite state space Ω_X, and suppose $P_X : \Omega_X \to [0, 1]$ is a probability mass function (PMF) for X, i.e., $P_X(x) \geq 0$ for all $x \in \Omega_X$, and $\sum_{x \in \Omega_X} P_X(x) = 1$. Then the expected value of X with respect to P_X, denoted by $E_{P_X}(X)$, is defined as follows:

$$E_{P_X}(X) = \sum_{x \in \Omega_X} x \cdot P_X(x) \tag{1}$$

Notice that as X is real-valued, the definition in Eq. (1) is well defined. Also, as Ω_X is finite, $E_{P_X}(X)$ always exists.

2.2 Properties of probabilistic expectation

Consider the situation in the definition of probabilistic expectation. The expectation operator has the following properties.

1. (*Expected value of a constant*) If X is a constant, i.e., $P_X(a) = 1$, where a is a real constant, then $E_{P_X}(X) = a$.

2. (*Expected value of a function of X*) Suppose $Y = g_X : \Omega_X \rightarrow \mathbb{R}$ is a well-defined function of X, where $\mathbb{R}$ is the set of all real numbers. Then, $E_{P_Y}(Y)$ is as follows:

$$E_{P_Y}(Y) = \sum_{x \in \Omega_X} g_X(x) \cdot P_X(x) \tag{2}$$

For convenience, the right-hand-side of Eq. (2) is denoted by $E_{P_X}(g_X)$. This property is referred to as the *law of the unconscious statistician*. If $Y = g_X$, then Y is a random variable whose PMF P_Y is defined in terms of PMF P_X as follows:

$$P_Y(y) = \sum_{x \in \Omega_X : g_X(x) = y} P_X(x) \tag{3}$$

It follows from the definition of expected value that $E_{P_Y}(Y) = E_{P_X}(g_X)$. The result in Eq. (2) says that $E_{P_Y}(Y)$ can be computed directly from the PMF of X without computing the PMF of Y.

3. (*Expected value of a linear function of X*) Suppose $Y = g_X = aX + b$ where a and b are real constants. Then $E_{P_Y}(Y) = aE_{P_X}(X) + b$.

4. (*Expected value of a function of X and Y*) The law of the unconscious statistician generalizes to the multidimensional case. Suppose X and Y are discrete random variables with state spaces Ω_X and Ω_Y, respectively, with joint PMF $P_{X,Y}$, i.e., $P_{X,Y}(x, y) \geq 0$ for all $(x, y) \in \Omega_X \times \Omega_Y$, and $\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} P_{X,Y}(x, y) = 1$. Then if $Z = g_{X,Y} : \Omega_X \times \Omega_Y \rightarrow \mathbb{R}$ is a well-defined function of (X, Y), then

$$E_{P_Z}(Z) = E_{P_{X,Y}}(g_{X,Y}) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} g_{X,Y}(x, y) P_{X,Y}(x, y) \tag{4}$$

5. (*Expected value of a linear function of X and Y*) If $Z = g_{X,Y} = aX + bY + c$, where a, b, and c are real constants, then

$$E_{P_Z}(Z) = aE_{P_{X,Y}}(X) + bE_{P_{X,Y}}(Y) + c \tag{5}$$

3 Basic Definitions in the D-S Belief Functions Theory

In this section, we review the basic definitions in the D-S belief functions theory. Like the various uncertainty theories, D-S belief functions theory includes functional representations of uncertain knowledge, and operations for making inferences from such knowledge. Most of this material is taken from [5].

3.1 Representations of belief functions

Belief functions can be represented in four different ways: basic probability assignments (BPA), belief functions, plausibility functions, and commonality functions. Here, we focus only on BPA and commonality functions.

Suppose X is a random variable with state space Ω_X. Let 2^{Ω_X} denote the set of all *non-empty* subsets of Ω_X. A basic probability assignment (BPA) m_X for X is a function $m_X : 2^{\Omega_X} \to [0, 1]$ such that $\sum_{\mathsf{a} \in 2^{\Omega_X}} m_X(\mathsf{a}) = 1$.

The non-empty subsets $\mathsf{a} \in 2^{\Omega_X}$ such that $m_X(\mathsf{a}) > 0$ are called *focal* elements of m_X. An example of a BPA for X is the vacuous BPA for X, denoted by ι_X, such that $\iota_X(\Omega_X) = 1$. If all focal elements of m_X are singleton subsets of Ω_X, then we say m_X is *Bayesian*. In this case, m_X is equivalent to the PMF P_X for X such that $P_X(x) = m_X(\{x\})$ for each $x \in \Omega_X$.

The information in a BPA m_X can also be represented by a corresponding commonality function Q_{m_X} that is defined as follows: $Q_{m_X}(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_X} : \mathsf{b} \supseteq \mathsf{a}} m_X(\mathsf{b})$ for all $\mathsf{a} \in 2^{\Omega_X}$. For the example above with $\Omega_X = \{x, \bar{x}\}$, the commonality function Q_{ι_X} corresponding to BPA ι_X is given by $Q_{\iota_X}(\{x\}) = 1$, $Q_{\iota_X}(\{\bar{x}\}) = 1$, and $Q_{\iota_X}(\Omega_X) = 1$. If m_X is a Bayesian BPA for X, then Q_{m_X} is such that $Q_{m_X}(\mathsf{a}) = m_X(\mathsf{a})$ if $|\mathsf{a}| = 1$, and $Q_m(\mathsf{a}) = 0$ if $|\mathsf{a}| > 1$. Q_{m_X} is a non-increasing function in the sense that if $\mathsf{b} \subseteq \mathsf{a}$, then $Q_{m_X}(\mathsf{b}) \geq Q_{m_X}(\mathsf{a})$. Finally, Q_{m_X} is a normalized function in the sense that:

$$\begin{aligned}
\sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} Q_{m_X}(\mathsf{a}) &= \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} \left(\sum_{\mathsf{b} \in 2^{\Omega_X} : \mathsf{b} \supseteq \mathsf{a}} m_X(\mathsf{b}) \right) \\
&= \sum_{\mathsf{b} \in 2^{\Omega_X}} m_X(\mathsf{b}) \left(\sum_{\mathsf{a} \in 2^{\Omega_X} : \mathsf{a} \subseteq \mathsf{b}} (-1)^{|\mathsf{a}|+1} \right) \\
&= \sum_{\mathsf{b} \in 2^{\Omega_X}} m_X(\mathsf{b}) = 1.
\end{aligned}$$

Next, we describe the two main operations for making inferences.

3.2 Basic operations in the D-S theory

There are two main operations in the D-S theory—Dempster's combination rule and marginalization.

In the D-S theory, we can combine two BPAs m_1 and m_2 representing distinct pieces of evidence by Dempster's rule [2] and obtain the BPA $m_1 \oplus m_2$, which represents the combined evidence. Dempster referred to this rule as the product-intersection rule, as the product of the BPA values are assigned to the intersection of the focal elements, followed by normalization. Normalization consists of discarding the probability assigned to $\emptyset$, and normalizing the remaining values so that they add to 1. In general, Dempster's rule of combination can be used to combine two BPAs for arbitrary sets of variables.

Let $\mathcal{X}$ denote a finite set of variables. The state space of $\mathcal{X}$ is $\times_{X\in\mathcal{X}}\Omega_X$. Thus, if $\mathcal{X} = \{X, Y\}$ then the state space of $\{X, Y\}$ is $\Omega_X \times \Omega_Y$.

Projection of states simply means dropping extra coordinates; for example, if (x, y) is a state of $\{X, Y\}$, then the projection of (x, y) to X, denoted by $(x, y)^{\downarrow X}$, is simply x, which is a state of X.

Projection of subsets of states is achieved by projecting every state in the subset. Suppose $\mathsf{b} \in 2^{\Omega_{\{X,Y\}}}$. Then $\mathsf{b}^{\downarrow X} = \{x \in \Omega_X : (x, y) \in \mathsf{b}\}$. Notice that $\mathsf{b}^{\downarrow X} \in 2^{\Omega_X}$.

Dempster's rule can be defined in terms of commonality functions [6] as follows: Suppose m_1 and m_2 are BPAs for $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Suppose Q_{m_1} and Q_{m_2} are commonality functions corresponding to BPAs m_1 and m_2, respectively. The commonality function $Q_{m_1\oplus m_2}$ corresponding to BPA $m_1 \oplus m_2$ for $\mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$ is as follows:

$$Q_{m_1\oplus m_2}(\mathsf{a}) = K^{-1} Q_{m_1}(\mathsf{a}^{\downarrow\mathcal{X}_1})\, Q_{m_2}(\mathsf{a}^{\downarrow\mathcal{X}_2}), \tag{6}$$

for all $\mathsf{a} \in 2^{\Omega_\mathcal{X}}$, where the normalization constant K is as follows:

$$K = \sum_{\mathsf{a}\in 2^{\Omega_\mathcal{X}}} (-1)^{|\mathsf{a}|+1} Q_{m_1}(\mathsf{a}^{\downarrow\mathcal{X}_1})\, Q_{m_2}(\mathsf{a}^{\downarrow\mathcal{X}_2}). \tag{7}$$

The definition of Dempster's rule assumes that the normalization constant K is non-zero. If $K = 0$, then the two BPAs m_1 and m_2 are said to be in *total conflict* and cannot be combined. In terms of commonality functions, Dempster's rule is pointwise multiplication of commonality functions followed by normalization.

Marginalization in D-S theory is addition of values of BPAs. Suppose m is a BPA for $\mathcal{X}$. Then, the marginal of m for $\mathcal{X}_1$, where $\mathcal{X}_1 \subseteq \mathcal{X}$, denoted by $m^{\downarrow\mathcal{X}_1}$, is a BPA for $\mathcal{X}_1$ such that for each $\mathsf{a} \in 2^{\Omega_{\mathcal{X}_1}}$,

$$m^{\downarrow\mathcal{X}_1}(\mathsf{a}) = \sum_{\mathsf{b}\in 2^{\Omega_\mathcal{X}} : \mathsf{b}^{\downarrow\mathcal{X}_1} = \mathsf{a}} m(\mathsf{b}). \tag{8}$$

This completes a brief description of D-S theory of belief functions. For more details, see [6].

4 A New Definition of Expected Value for the D-S Theory

In this section, we provide a new definition of expected value of belief functions in the D-S theory, and describe its properties.

As in the probabilistic case, we will assume that Ω_X is a finite set of real numbers. In a PMF, we have probabilities assigned to each state $x \in \Omega_X$. In a BPA m_X for X and its equivalent representations, we have probabilities assigned to subsets of states $\mathsf{a} \in 2^{\Omega_X}$. Before we define expected value of X with respect to BPA m_X, we will define a real-valued value function $v_{m_X} : 2^{\Omega_X} \to \mathbb{R}$ for all subsets in 2^{Ω_X}. If $\mathsf{a} = \{x\}$ is a singleton subset, then we can consider $v_m(\{x\}) = x$.

Remember that the elements of Ω_X are real numbers. For non-singleton subsets $\mathsf{a} \in 2^{\Omega_X}$, it makes sense to define $v_{m_X}(\mathsf{a})$ such that the following inequality holds:

$$\min \mathsf{a} \leq v_{m_X}(\mathsf{a}) \leq \max \mathsf{a} \tag{9}$$

One way to satisfy the inequality in Eq. (9) is as follows:

$$v_{m_X}(\mathsf{a}) = \frac{\sum_{x \in \mathsf{a}} x \cdot Q_{m_X}(\{x\})}{\sum_{x \in \mathsf{a}} Q_{m_X}(\{x\})} \quad \text{for all } \mathsf{a} \in 2^{\Omega_X} \tag{10}$$

In words, the value function $v_{m_X}(\mathsf{a})$ is the weighted average of all $x \in \mathsf{a}$, where the weights are the commonality numbers $Q_{m_X}(\{x\})$.

4.1 Definition of expected value for D-S belief functions

Suppose m_X is a BPA for X with real-valued state space Ω_X, and suppose Q_{m_X} denotes the commonality function corresponding to m_X. Then the expected value of X with respect to m_X, denoted by $E_{m_X}(X)$, is defined as follows:

$$E_{m_X}(X) = \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} v_{m_X}(\mathsf{a}) \cdot Q_{m_X}(\mathsf{a}) \tag{11}$$

4.2 Properties of expected values of D-S belief functions

Some important properties of our definition in Eq. (11) are as follows. Consider the situation in the definition of expected value of D-S belief functions in Eq. (11).

1. (*Consistency with probabilistic expectation*) If m_X is a Bayesian BPA for X, and P_X is the PMF for X corresponding to m_X, i.e., $P_X(x) = m_X(\{x\})$ for all $x \in \Omega_X$, then $E_{m_X}(X) = E_{P_X}(X)$.

   Proof: As m_X is Bayesian, $Q_{m_X}(\mathsf{a}) = m_X(\mathsf{a})$ if $|\mathsf{a}| = 1$, and $Q_{m_X}(\mathsf{a}) = 0$ if $|\mathsf{a}| > 1$. Also, $v_{m_X}(\{x\}) = x$. Thus, $E_{m_X}(X)$ in Eq. (11) reduces to $E_{P_X}(X)$ in Eq. (1).

2. (*Expectation of a constant*) If X is a constant, i.e., $m_X(\{a\}) = 1$, where a is a real constant, then $E_{m_X}(X) = a$.

   Proof: Notice that in this case, m is Bayesian, and as this property holds for the probabilistic case, it also holds for the D-S theory from the *consistency with probabilistic expectation* property.

3. (*Expected value of a function of X*) Suppose $Y = g_X : \Omega_X \to \mathbb{R}$ is a *linear* function, then $E_{m_Y}(Y)$ can be computed as follows:

$$E_{m_Y}(Y) = E_{m_X}(g_X) = \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} \, g_X(v_{m_X}(\mathsf{a})) \, Q_{m_X}(\mathsf{a}) \tag{12}$$

In probability theory, this property is valid for any well-defined function of X. Our definition does not satisfy this property for any well-defined function (see Examples 1 and 2 that follow), but it is satisfied only for a linear function of X. This property allows us to compute the expected value of $Y = g_X$ without first computing its commonality function.

Proof: As g_X is linear, it is a 1-1 function. Therefore, $\Omega_Y = \{g_X(x) : x \in \Omega_X\}$. Thus, the values of the commonality function Q_{m_Y} for Y are the same as the corresponding values of the commonality function Q_m for X, i.e., $Q_{m_Y}(\mathsf{a_Y}) = Q_{m_X}(\mathsf{a})$, where $\mathsf{a_Y} \in 2^{\Omega_Y}$ is the subset that corresponds to subset a of Ω_X, i.e., $\mathsf{a}_Y = \{g_X(x) : x \in \mathsf{a}\}$. It suffices to show that $v_{m_Y}(\mathsf{a_Y}) = g(v_m(\mathsf{a}))$ for all $\mathsf{a} \in 2^{\Omega_X}$. Suppose $Y = g_X = aX + b$.

$$\begin{aligned} v_{m_Y}(\mathsf{a}_Y) &= \frac{\sum_{y \in \mathsf{a}_Y} y \cdot Q_{m_Y}(\{y\})}{\sum_{y \in \mathsf{a}_Y} Q_{m_Y}(\{y\})} \\ &= \frac{\sum_{x \in \mathsf{a}} (ax + b) \cdot Q_{m_X}(\{x\})}{\sum_{x \in \mathsf{a}} Q_{m_X}(\{x\})} \\ &= a \frac{\sum_{x \in \mathsf{a}} x \cdot Q_{m_X}(\{x\})}{\sum_{x \in \mathsf{a}} Q_{m_X}(\{x\})} + b \\ &= a v_{m_X}(\mathsf{a}) + b \\ &= g_X(v_m(\mathsf{a})) \end{aligned}$$

This completes the proof.

4. (Expected value of a linear function of X) Suppose $Y = g_X = aX + b$ where a and b are real constants, and suppose m_X is a BPA for X. Then $E_{m_Y}(Y) = aE_{m_X}(X) + b$.

   Proof: From the *expected value of a function of X* property, it follows that that $E_{m_Y}(Y) = E_{m_X}(g_X) = E_{m_X}(aX + b)$. Thus,

$$\begin{aligned} E_{m_Y}(Y) &= \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} (a v_{m_X}(\mathsf{a}) + b) Q_{m_X}(\mathsf{a}) \\ &= a \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} v_{m_X}(\mathsf{a}) Q_{m_X}(\mathsf{a}) + b \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} Q_{m_X}(\mathsf{a}) \\ &= a E_{m_X}(X) + b. \end{aligned}$$

5. (*Expected value of a function of X and Y*) The law of the unconscious statistician generalizes to the multidimensional case. Suppose X and Y are discrete random variables with state spaces Ω_X and Ω_Y, respectively, with joint BPA $m_{X,Y}$ for (X, Y). If $g_{X,Y} : \Omega_X \times \Omega_Y \to \mathbb{R}$ is a *linear* function of (X, Y), then

$$E_m(g_{X,Y}) = \sum_{\mathsf{a} \in 2^{\Omega_X \times \Omega_Y}} (-1)^{|\mathsf{a}|+1} \, g_{X,Y}(v(\mathsf{a})) \, Q(\mathsf{a}) \tag{13}$$

Table 1: Expected value of a function $Y = g_X = X^2$ that is not 1-1

$\mathsf{a} \in 2^{\Omega_X}$	$m_X(\mathsf{a})$	$Q_{m_X}(\mathsf{a})$	$v_{m_X}(\mathsf{a})$	$E_{m_X}(X)$	$(v_{m_X}(\mathsf{a}))^2$	$E_{m_X}(g_X)$
$\{-1\}$	0.02	0.63	-1.00	0.059	1.00	1.188
$\{0\}$	0.05	0.70	0.00		0.00	
$\{1\}$	0.09	0.81	1.00		1.00	
$\{-1, 0\}$	0.12	0.42	-0.47		0.22	
$\{-1, 1\}$	0.19	0.49	0.13		0.02	
$\{0, 1\}$	0.23	0.53	0.54		0.29	
$\{-1, 0, 1\}$	0.30	0.30	0.08		0.01	
$\mathsf{b} \in 2^{\Omega_Y}$	$m_Y(\mathsf{b})$	$Q_{m_Y}(\mathsf{b})$	$v_{m_Y}(\mathsf{b})$	$E_{m_Y}(Y)$		
$\{1\}$	0.30	0.95	1.00	0.576		
$\{0\}$	0.05	0.70	0.00			
$\{1, 0\}$	0.65	0.65	0.58			

As in the case of *expected value of a function of X* property, this property holds only for the case where $g_{X,Y}$ is a linear function.

A proof of this property is similar to the proof of the *expected value of a function of X* property, and is therefore omitted.

6. (*Expected value of a linear function of X and Y*) If $Z = g_{X,Y} = aX + bY + c$, where a, b, and c are real constants, and $m_{X,Y}$ is a joint BPA for (X, Y), then

$$E_{m_Z}(Z) = E_{m_{X,Y}}(aX + bY + c) = aE_{m_{X,Y}}(X) + bE_{m_{X,Y}}(Y) + c \qquad (14)$$

A proof of this property is similar to the proof of the *expected value of a linear function of X* property, and is therefore omitted.

Example 1 (Non 1-1 function) *Consider a real-valued variable X with $\Omega_X = \{-1, 0, 1\}$, and suppose m_X is a BPA for X as shown in Table 1. Suppose $Y = g_X = X^2$. Notice that g_X is not 1-1. Then, $\Omega_Y = \{1, 0\}$, and m_Y is as shown in Table 1. For this example, $E_{m_Y}(Y) = 0.576$, and $E_{m_X}(g_X) = 1.188$. Thus, Eq.* (12) *does not hold.*

Example 2 (Nonlinear 1-1 function) *Consider a real-valued variable Z with $\Omega_X = \{1, 2, 3\}$, and suppose m_Z is a BPA for Z as shown in Table 2. Suppose $Y = g_Z = \log(Z)$. Then, $\Omega_Y = \{\log(1), \log(2), \log(3)\} \approx \{0, 0.30, 0.48\}$, and m_Y is as shown in Table 2. As the function is 1-1, the values of m_Y are the same as the values of m_Z. For this example, $E_{m_Y}(Y) = 0.273$, and $E_{m_Z}(\log(Z)) = 0.241$. Thus, Eq.* (12) *does not hold.*

Table 2: Expected value of $Y = g_Z = \log(Z)$, a nonlinear 1-1 function

$\mathsf{a} \in 2^{\Omega_Z}$	$m_Z(\mathsf{a})$	$Q_{m_Z}(\mathsf{a})$	$v_{m_Z}(\mathsf{a})$	$E_{m_Z}(Z)$	$\log(v_{m_Z}(\mathsf{a}))$	$E_{m_Z}(g_Z)$
$\{1\}$	0.02	0.63	1.00	2.059	0.00	0.241
$\{2\}$	0.05	0.70	2.00		0.30	
$\{3\}$	0.09	0.81	3.00		0.48	
$\{1,2\}$	0.12	0.42	1.53		0.18	
$\{1,3\}$	0.19	0.49	2.12		0.33	
$\{2,3\}$	0.23	0.53	2.53		0.40	
$\{1,2,3\}$	0.30	0.30	2.08		0.32	
$\mathsf{a}_Y \in 2^{\Omega_Y}$	$m_Y(\mathsf{a}_Y)$	$Q_{m_Y}(\mathsf{a}_Y)$	$v_{m_Y}(\mathsf{a}_Y)$	$E_{m_Y}(Y)$		
$\{0\}$	0.02	0.63	0.00	0.273		
$\{0.30\}$	0.05	0.70	0.30			
$\{0.48\}$	0.09	0.81	0.48			
$\{0, 0.30\}$	0.12	0.42	0.16			
$\{0, 0.48\}$	0.19	0.49	0.27			
$\{0.30, 048\}$	0.23	0.53	0.40			
$\{0, 0.30, 0.48\}$	0.30	0.30	0.28			

Example 3 (Linear function) *Consider a real-valued variable X with $\Omega_X = \{-1, 0, 1\}$, and suppose m_X is a BPA for X as shown in Table 3. Suppose $Y = g_X = 2X + 1$. Then, $\Omega_Y = \{-1, 1, 3\}$, and m_Y is as shown in Table 3. Notice that as a linear function is 1-1, the values of m_Y are the same as the corresponding values of m_X. Also, notice that as the function g_X is linear, $g(v_m(\mathsf{a})) = v_{m_Y}(\mathsf{a}_Y)$, where subset a_Y corresponds to subset a. For this example, $E_{m_Y}(Y) = 1.117$, and $E_{m_X}(g_X) = 1.117$. Thus, Eq. (12) holds. Also notice that $E_{m_X}(g_X) = E_{m_X}(2X + 1) = 2E_{m_X}(X) + 1 = 2(0.059) + 1 = 1.117$.*

4.3 A definition of expected value of a real-valued function of X

Suppose Q_{m_X} is a commonality function for X corresponding to BPA m_X for X, and Ω_X may not be real-valued, but $g_X : 2^{\Omega_X} \to \mathbb{R}$ is a well-defined real-valued function of X, then we define expected value of g_X with respect to m_X, denoted by $E_{m_X}(g_X)$, as follows:

$$E_{m_X}(g_X) = \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} \, g_X(\mathsf{a}) \, Q_{m_X}(\mathsf{a}) \tag{15}$$

The definition of the expected value of g_X with respect to m_X is similar to Eqs. (12) and (13). Such a definition may be useful in comparing preference for lotteries that are characterized by D-S belief functions similar to von Neumann-Morgenstern's utility theory for probabilistic lotteries [11].

Table 3: Expected value of $Y = 2X + 1$, a linear function

$\mathsf{a} \in 2^{\Omega_X}$	$m_X(\mathsf{a})$	$Q_{m_X}(\mathsf{a})$	$v_{m_X}(\mathsf{a})$	$E_{m_X}(X)$	$2\,v_{m_X}(\mathsf{a})) + 1$	$E_{m_X}(g_X)$
$\{-1\}$	0.02	0.63	-1.00	0.059	-1.00	1.117
$\{0\}$	0.05	0.70	0.00		1.00	
$\{1\}$	0.09	0.81	1.00		3.00	
$\{-1, 0\}$	0.12	0.42	-0.47		0.05	
$\{-1, 1\}$	0.19	0.49	0.12		1.24	
$\{0, 1\}$	0.23	0.53	0.53		2.07	
$\{-1, 0, 1\}$	0.30	0.30	0.08		1.16	

$\mathsf{a}_Y \in 2^{\Omega_Y}$	$m_Y(\mathsf{a}_Y)$	$Q_{m_Y}(\mathsf{a}_Y)$	$v_{m_Y}(\mathsf{a}_Y)$	$E_{m_Y}(Y)$
$\{-1\}$	0.02	0.63	-1.00	1.117
$\{1\}$	0.05	0.70	1.00	
$\{3\}$	0.09	0.81	3.00	
$\{-1, 1\}$	0.12	0.42	0.05	
$\{-1, 3\}$	0.19	0.49	1.24	
$\{1, 3\}$	0.23	0.53	2.07	
$\{-1, 1, 3\}$	0.30	0.30	1.16	

4.4 A comparison with expectation of pignistic and plausibility transforms

As we said earlier, a traditional method of computing expectations of random variables characterized by a D-S BPA is to first transform the BPA to a PMF, and then use the probabilistic expectation operator. There are several methods of transforming a BPA to a PMF. Here we focus on the pignistic [9] and the plausibility [1] transforms.

As D-S theory is a generalization of probability theory, there is, in general, more information in a BPA m than in the corresponding transform of m to a PMF. Thus, by computing expectation of X whose uncertainty is described by BPA m by first transforming m to a pignistic PMF $BetP_m$, or to a plausibility PMF Pl_P_m, there may be loss of information.

In general, the expected value defined in this paper may yield different values than the probabilistic expectation using pignistic or plausibility transformation. Table 4 compares the expectation defined in this paper with probabilisitic expectation using pignistic and plausibility transforms for the various BPAs described in Tables 1, 2, and 3. Two observations. First, although the three definitions yield different answers, they are all approximately of the same order of magnitude. Second, all three definitions satisfy the *expected value of a linear function of X* property. Thus, BPA m_Z in Table 2 can be obtained from BPA m_X in Table 1 using the transformation $Z = X + 2$. All three expected values satisfy the *expected value of a linear function of X* property. Also, BPA m_Y in Table 3 is obtained from BPA m_X in Table 1 using the transformation $Y = 2X + 1$. Again, all three expected values satisfy the *expected value of a linear function of X* property.

Table 4: A comparison of our expected value with probabilistic expectation using pignistic and plausibility transforms

BPA m	$E_m(\cdot)$	$E_{BetP_m}(\cdot)$	$E_{Pl\_P_m}(\cdot)$
m_X in Table 1	0.059	0.125	0.084
m_Y in Table 1	0.576	0.625	0.576
m_Z in Table 2	2.059	2.125	2.084
m_Y in Table 2	0.273	0.289	0.278
m_Y in Table 3	1.117	1.250	1.168

5 Summary and Conclusions

We propose a new definition of expected value for real-valued random variables whose uncertainty is described by D-S belief functions. Also, if we have a random variable with a symbolic frame of discernment, but a real-valued function defined on the set of all non-empty subsets of the frame, then we propose a new definition of expectation of the function in a similar manner.

Our new definition satisfies many of the properties satisfied by the probabilistic expectation operator, which was first proposed by Christiaan Huygens [3] in the context of the problem of points posed by Chevalier de Méré to Blaise Pascal.

The expectation operator can be used to define variance, covariance, correlation, and higher moments of D-S belief functions [8].

If we define $I(\mathsf{a}) = \log_2(\frac{1}{Q_{m_X}(\mathsf{a})})$ as the information content of observing subset $\mathsf{a} \in 2^{\Omega_X}$ whose uncertainty is described by m_X, then similar to Shannon's definition of entropy of PMFs [7], we define entropy of BPA m_X for X as an expected value of the function $I(\mathsf{a})$, i.e., $H(m_X) = E_{m_X}(I(\mathsf{a}))$. This is what is proposed in [5]. This definition of entropy has many nice properties. In particular, it satisfies the compound distributions property: $H(m_X \oplus m_{Y|X}) = H(m_X) + H(m_{Y|X})$, where $m_{Y|X}$ is a conditional BPA for Y given X obtained by $\oplus\{m_{Y|x} : x \in \Omega_X\}$, and $m_{Y|x}$ is a conditional BPA for Y given $X = x$.

There are several decision theories for lotteries whose uncertainty is described by D-S belief functions theory. The most prominent ones are by Jean-Yves Jaffray [4]/Thomas Strat [10], and Philippe Smets [9]. The proposal by Jaffray/Strat is to first reduce a D-S belief function to an upper and lower PMFs, and then define an expected value that is a convex combination of the upper and lower probabilistic expectation. This proposal is justified in [4] by some axioms similar to the axioms proposed by John von Neumann and Oskar Morgenstern [11] for probabilistic lotteries. The proposal by Smets is to transform a D-S belief function to a corresponding PMF called the pignistic transform, and then use von Neumann-Morgenstern's expected utility theory. Our definition of expected value can be used in a decision theory for D-S theory without transforming belief functions to PMFs. This remains to be done.

Acknowledgments

This paper is a short version of [8], which is inspired by recent work on entropy of belief functions in the Dempster-Shafer theory of evidence in collaboration with Radim Jiroušek from The Czech Academy of Sciences [5]. Thank you, Radim! Thanks also to Suzanna Emelio for a careful proofreading of this paper.

References

[1] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–340, 2006.

[2] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[3] C. Huygens. *De ratiociniis in ludo aleæ*. 1657 (English translation, published in 1714).

[4] J.-Y. Jaffray. Linear utility theory for belief functons. *Operations Research Letters*, 8(2):107–112, 1989.

[5] R. Jiroušek and P. P. Shenoy. A decomposable entropy of belief functions in the Dempster-Shafer theory. Working Paper 334, University of Kansas School of Business, Lawrence, KS 66045 USA, 2018.

[6] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[7] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[8] P. P. Shenoy. An expectation operator for belief functions in the Dempster-Shafer theory. Working Paper 335, University of Kansas School of Business, Lawrence, KS 66045, 2018.

[9] P. Smets. Decision making in a context where uncertainty is represented by belief functions. In R. P. Srivastava and T. J. Mock, editors, *Belief Functions in Business Decisions*, pages 316–332. Physica-Verlag, Heidelberg, 2002.

[10] T. M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4(5–6):391–417, 1990.

[11] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Priceton, NJ, 1953.

On attempts to characterize facet-defining inequalities of the cone of exact games

Milan Studený studeny@utia.cas.cz
Tomáš Kroupa kroupa@utia.cas.cz
Václav Kratochvíl velorex@utia.cas.cz

Department of Decision Making Theory
The Czech Academy of Sciences, Institute of Information Theory and Automation

Abstract

The sets of balanced, totally balanced, exact and supermodular games play an important role in cooperative game theory. These sets of games are known to be polyhedral cones. The (unique) non-redundant description of these cones by means of the so-called facet-defining inequalities is known in cases of balanced games and supermodular games, respectively. The facet description of the cones of exact games and totally balanced games are not known and we present conjectures about what are the facet-defining inequalities for these cones.

We introduce the concept of an *irreducible min-balanced* set system and conjecture that the facet-defining inequalities for the cone of totally balanced games correspond to these set systems. The conjecture concerning exact games is that the facet-defining inequalities for this cone are those which correspond to irreducible min-balanced systems on strict subsets of the set of players and their conjugate inequalities. A consequence of the validity of the conjectures would be a novel result saying that a game m is exact if and only if m and its reflection are totally balanced.

1 Introduction: former results overview

Important classes of set functions used as mathematical models in the coalition game theory are: the class of *balanced games* $\mathcal{B}(N)$, the class of *totally balanced games* $\mathcal{T}(N)$, the class of *exact games* $\mathcal{E}(N)$ and the class of *supermodular games* $\mathcal{S}(N)$, named traditionally *convex* in game-theoretical community. One has

$$\mathcal{B}(N) \supseteq \mathcal{T}(N) \supseteq \mathcal{E}(N) \supseteq \mathcal{S}(N)$$

and it is well-known that all these sets are polyhedral cones in the space $\mathbb{R}^{\mathcal{P}(N)}$, where $\mathcal{P}(N) = \{A : A \subseteq N\}$ is the power set of the set of players N. That means

that each of the cones can be specified by finitely many linear inequalities.

These set functions occur in other contexts, for example, in the context of *imprecise probabilities*. More specifically, supermodular games correspond to 2-monotone lower probabilities, exact games to coherent lower probabilities and balanced games to lower probabilities avoiding sure loss [9].

Recall that every full-dimensional polyhedral cone K in an Euclidean space has uniquely determined set of the so-called *facet-defining* inequalities, where the uniqueness of each inequality is up to a positive multiple. Specifically, these inequalities determine proper faces of K of maximal dimension, called *facets*. The complete list of facet-defining inequalities then provides the least possible inequality description of K, unique up to positive multiples.

The above cones $\mathcal{B}(N), \ldots, \mathcal{S}(N)$ are not full-dimensional in $\mathbb{R}^{\mathcal{P}(N)}$ but adding a one-dimensional linear space $\mathcal{C}(N)$ of constant functions turns them into full-dimensional cones $B(N), \ldots, S(N)$. The facet-defining inequalities of these extended cones then induce a non-redundant inequality description of the original cones of games. The non-redundant inequality description of these cones is known in cases of supermodular and balanced games only.

The facets of the *supermodular* cone $S(N)$ are defined by inequalities of the form $m(\{i,j\} \cup L) + m(L) - m(\{i\} \cup L) - m(\{j\} \cup L) \geq 0$ for $m \in S(N)$, where $L \subset N$ and $i, j \in N \setminus L$ are distinct [6]; these inequalities were known to correspond to elementary conditional independence statements [13].

The non-redundant inequalities for the cone $\mathcal{B}(N)$ of *balanced* games were characterized by Shapley [12] on basis of former results by Bondareva [1]. These inequalities correspond to "minimal balanced collections" of subsets of N whose union is N; in this paper we call such collections *min-balanced systems* on N.

The consequence of the fact that $\mathcal{B}(N)$ is a polyhedral cone is the observation that the set of totally balanced games $\mathcal{T}(N)$ is a polyhedral cone. Nonetheless, as far as we know, the facet-defining inequalities for the cone $\mathcal{T}(N)$ have not been described/discussed in the literature.

Recently, the fact that set of exact games $\mathcal{E}(N)$ forms a convex cone has been derived [3]. Shortly after that Lohman *et al.* [8] even showed that the set of exact games is a polyhedral cone. Specifically, the exact games were characterized by means of finitely many linear inequalities that correspond to the so-called "minimal exact balanced" collections of subsets of N. Although finitely many linear inequalities specifying $\mathcal{E}(N)$ were classified in [8], many of these inequalities are already known to be redundant.

In this paper we mainly deal with the question of what are the facet-defining inequalities for the exact cone. On basis of our own computations as well as computations made by Quaeghebeur in connection with his thesis [11] we found and classified these inequalities in case $|N| \leq 5$. We analyzed the results and revealed certain symmetry in the problem. More specifically, the facet-defining inequalities for $E(N)$ come in pairs: every such inequality is accompanied with a *conjugate* one. We have shown that this is a consequence of the fact that the cone $E(N)$ is closed under a special *reflection* transformation.

We even came to sensible conjectures about what are the facet-defining inequalities for the cones $T(N)$ and $E(N)$. The basis of them is the concept of a *min-balanced system* on $M \subseteq N$, where $|M| \geq 2$, which is a certain collection of subsets of M. Special *irreducible min-balanced systems* seem to play the crucial role. The conjecture concerning the totally balanced cone is that these irreducible systems correspond to facets of $T(N)$. The conjecture concerning the exact cone is that every facet-defining inequality for $E(N)$ is either given by an irreducible min-balanced system on some strict subset $M \subset N$, $|M| \geq 2$, or it a conjugate inequality to such an inequality.

We also briefly report on our effort to develop a computer programme for generating all (permutational types of) min-balanced systems and irreducible min-balanced systems. We employed the algorithm by Peleg [10] and reformulated the problem in terms special bipartite graphs using the BLISS algorithm by Junttila and Kaski [5].

2 Basic concepts and facts

Throughout the text N is a finite set of *players* such that $|N| \geq 2$. Given $S \subseteq N$, the symbol χ_S will denote zero-one incidence vector of S (in $\mathbb{R}^N$); that is, $\chi_S(i) = 1$ if $i \in S$ and $\chi_S(j) = 0$ if $j \in N \setminus S$.

A *game* is a set function $m : \mathcal{P}(N) \to \mathbb{R}$ such that $m(\emptyset) = 0$. The *core* of a game m is a polytope (= bounded polyhedron) in $\mathbb{R}^N$ given by

$$C(m) \;:=\; \{\, [x_i]_{i\in N} \;:\; \sum_{i\in N} x_i = m(N) \quad \& \quad \forall\, S \subseteq N \quad \sum_{i\in S} x_i \geq m(S) \,\}\,.$$

A game m is *balanced* if it has a non-empty core: $C(m) \neq \emptyset$; it is called *totally balanced* if, for each $M \subseteq N$, $|M| \geq 2$, the restriction of m to $\mathcal{P}(M)$ is balanced. A balanced game is called *exact* if every lower bound is tight:

$$\forall\, S \subseteq N \quad \exists\, x \in C(m) \qquad \sum_{i\in S} x_i = m(S)\,;$$

an equivalent definition is that m can be reconstructed from its core by minimization: $m(S) = \min\{\, \sum_{i\in S} x_i \;:\; x \in C(m)\}$ for any $S \subseteq N$. Well-known facts are that every exact game is totally balanced and that every supermodular game is exact [4].

Definition 1 A system $\mathcal{B} \subseteq \mathcal{P}(N)$ is *min-balanced on* a non-empty set $M \subseteq N$ if it is a minimal set system such that χ_M is in the conic hull of $\{\chi_S \;:\; S \in \mathcal{B}\}$.

Of course, the minimality is meant in sense of inclusion of set systems.

Lemma 2 A set system $\mathcal{B} \subseteq \mathcal{P}(N)$ is min-balanced on $\emptyset \neq M \subseteq N$ if and only if the following two conditions hold:

(i) there exist strictly positive coefficients $\lambda_S > 0$ for $S \in \mathcal{B}$ such that

$$\chi_M = \sum_{S \in \mathcal{B}} \lambda_S \cdot \chi_S \qquad \text{where } M = \bigcup \mathcal{B}, \text{ and}$$

(ii) the incidence vectors $\{\chi_S \in \mathbb{R}^N : S \in \mathcal{B}\}$ are linearly independent.

Hence, $\mathcal{B} \subseteq \mathcal{P}(N)$ is min-balanced iff it is a minimal set system satisfying (i).

The condition (i) is the *balancedness* condition from [12]; (ii) is equivalent to minimality and implies the uniqueness of the coefficients λ_S in (i).

Proof. To show the necessity of (i) write $\chi_M = \sum_{S \in \mathcal{B}} \lambda_S \cdot \chi_S$ with $\lambda_S \geq 0$. If λ_S vanishes for some S then we take $\mathcal{B}' = \{T \in \mathcal{B} : \lambda_T > 0\}$ to get a contradictory conclusion that $\mathcal{B}'$ is a strict subsystem of $\mathcal{B}$ satisfying the requirement. The necessity of (ii) can then be shown by a contradiction: otherwise a non-vanishing system of coefficients $\{\gamma_S : S \in \mathcal{B}\}$ exists such that $\sum_{S \in \mathcal{B}} \gamma_S \cdot \chi_S = \mathsf{0} \in \mathbb{R}^N$. For any $\varepsilon \geq 0$ put $\lambda^\varepsilon_S := \lambda_S + \varepsilon \cdot \gamma_S$ and consider $\chi_M = \sum_{S \in \mathcal{B}} \lambda^\varepsilon_S \cdot \chi_S$. Since all λ_S are strictly positive, maximal $\varepsilon > 0$ exists such that λ^ε_S are all non-negative. Put $\mathcal{B}' = \{T \in \mathcal{B} : \lambda^\varepsilon_T > 0\}$ and derive the contradiction analogously.

Conversely, if both (i) and (ii) holds then $\chi_M = \sum_{S \in \mathcal{B}} \lambda_S \cdot \chi_S$ with $\lambda_S > 0$. Assume for a contradiction that $\mathcal{C} \subset \mathcal{B}$ exists such that $\chi_M = \sum_{S \in \mathcal{C}} \nu_S \cdot \chi_S$ with $\nu_S \geq 0$, $S \in \mathcal{C}$. Put $\nu_S = 0$ for $S \in \mathcal{B} \setminus \mathcal{C}$. Then $\mathsf{0} = \sum_{S \in \mathcal{B}} (\lambda_S - \nu_S) \cdot \chi_S$, which contradicts (ii). The last claim is easy to derive from the former one. □

We intentionally restrict our attention to *non-trivial* min-balanced systems $\mathcal{B}$ with $|\mathcal{B}| \geq 2$; each such a system is ascribed the following inequality

$$m(\bigcup \mathcal{B}) - \sum_{S \in \mathcal{B}} \lambda_S \cdot m(S) \geq 0 \tag{1}$$

in which variables are represented by $m(S)$, $S \subseteq N$. We have shown in [7, Observation 4] that any non-trivial min-balanced system $\mathcal{B}$ on M satisfies the following conditions:

- the intersection $\bigcap \mathcal{B}$ is empty, one has $\emptyset, M \notin \mathcal{B}$, $|M| \geq 2$, and
- there are at most $|M|$ sets in $\mathcal{B}$.

The result from [12] is as follows.

Proposition 3 *The facet-defining inequalities for the cone $\mathcal{B}(N)$ are just the inequalities* (1) *for non-trivial min-balanced systems $\mathcal{B}$ on N.*

3 Conjugate inequalities

To reveal some important symmetry in the problem, it is suitable to consider the space $\mathbb{R}^{\mathcal{P}(N)}$ and extend all considered cones $\mathcal{B}(N), \ldots, \mathcal{S}(N)$ to this space. Formally, for $m \in \mathbb{R}^{\mathcal{P}(N)}$, a shifted function $\widetilde{m}$ given by $\widetilde{m}(S) := m(S) - m(\emptyset)$ for $S \subseteq N$ is a game and one can define:

$$\begin{aligned} B(N) &:= \{\, m \in \mathbb{R}^{\mathcal{P}(N)} \,:\, \widetilde{m} \text{ is a balanced game} \,\}, \\ T(N) &:= \{\, m \in \mathbb{R}^{\mathcal{P}(N)} \,:\, \widetilde{m} \text{ is a totally balanced game} \,\}, \\ E(N) &:= \{\, m \in \mathbb{R}^{\mathcal{P}(N)} \,:\, \widetilde{m} \text{ is an exact game} \,\}, \\ S(N) &:= \{\, m \in \mathbb{R}^{\mathcal{P}(N)} \,:\, \widetilde{m} \text{ is a supermodular game} \,\}. \end{aligned}$$

All these cones are full-dimensional in $\mathbb{R}^{\mathcal{P}(N)}$ and their shared linearity space appears to be the linear space of *modular* functions

$$L(N) := \{m \in \mathbb{R}^{\mathcal{P}(N)} : \, m(C \cup D) + m(C \cap D) = m(C) + m(D) \;\text{ for } C, D \subseteq N\},$$

which has the dimension $1 + |N|$; see [7, § 1.1].

The task to find/characterize facets of the original cones $\mathcal{B}(N), \ldots, \mathcal{S}(N)$ of games appears to be equivalent to finding facets of the above extended cones. Some geometric considerations lead to the conclusion that every facet-defining inequality for such a full-dimensional cone K in $\mathbb{R}^{\mathcal{P}(N)}$ has the form

$$\sum_{S \subseteq N} \alpha(S) \cdot m(S) \geq 0 \qquad \text{for } m \in \mathbb{R}^{\mathcal{P}(N)}, \tag{2}$$
$$\text{where } \textstyle\sum_{S \subseteq N} \alpha(S) = 0, \, \sum_{S \subseteq N : i \in S} \alpha(S) = 0 \text{ for any } i \in N,$$

and the coefficients $\alpha(S)$, $S \subseteq N$, are rational numbers. Thus, without loss of generality, we can multiply (2) by a positive factor to obtain relatively prime integers as coefficients. This is a *standardized form* of the inequality (2).

An inequality of the form (2) for $m \in B(N), \ldots, S(N)$ can be identified with the inequality $\sum_{\emptyset \neq S \subseteq N} \alpha(S) \cdot \widetilde{m}(S) \geq 0$ for $\widetilde{m} \in \mathcal{B}(N), \ldots, \mathcal{S}(N)$: for the inverse relation put $\alpha(\emptyset) := -\sum_{\emptyset \neq S \subseteq N} \alpha(S)$; for details see [7, § 1.1]. In particular, the inequality (1) has an extended version

$$m(\bigcup \mathcal{B}) - \sum_{S \in \mathcal{B}} \lambda_S \cdot m(S) + (-1 + \sum_{S \in \mathcal{B}} \lambda_S) \cdot m(\emptyset) \geq 0\,.$$

The point is that the considered cones, except for $T(N)$, are closed under the following linear self-transformation of $\mathbb{R}^{\mathcal{P}(N)}$. By a *reflection* of $m \in \mathbb{R}^{\mathcal{P}(N)}$ we mean $m^* \in \mathbb{R}^{\mathcal{P}(N)}$ given by

$$m^*(T) := m(N \setminus T) \qquad \text{for any } T \subseteq N.$$

It is nothing but inner composition with the "complement" mapping.

The inequality (2) can be ascribed a *conjugate inequality* of the form

$$\sum_{T\subseteq N} \alpha^*(T)\cdot m(T) \geq 0 \qquad \text{where } \alpha^*(T) := \alpha(N\setminus T) \text{ for any } T\subseteq N, \tag{3}$$

required for $m \in \mathbb{R}^{\mathcal{P}(N)}$. An important observation appears to be the equality

$$\sum_{T\subseteq N} \alpha^*(T)\cdot m(T) = \sum_{T\subseteq N} \alpha(N\setminus T)\cdot m^*(N\setminus T) = \sum_{S\subseteq N} \alpha(S)\cdot m^*(S)\,,$$

which easily implies that, whenever (2) is valid for vectors in a cone K which is closed under reflection, then (3) is valid for vector in K, and, of course, vice versa. In fact, one of our theoretical results is that (2) is facet-defining for K closed under reflection iff (3) is facet-defining for K [7, Lemma 26].

Every inequality of the form (2) defines a set system

$$\mathcal{B}_\alpha \;:=\; \{S\subseteq N \,:\, \alpha(S)<0\}\,. \tag{4}$$

Our analysis of facet-defining inequalities for $E(N)$ in case $|N|\leq 5$ written in the form (2) revealed that every system $\mathcal{B}_\alpha$ is either min-balanced on $M\subset N$ or it is *conjugate* to such a system $\mathcal{B}$, which means it is of the form

$$\mathcal{B}^* \;:=\; \{N\setminus S \,:\, S\in\mathcal{B}\}\,.$$

We explain now that any min-balanced system $\mathcal{B}$ defines a unique standardized inequality (2) with $\alpha\in\mathbb{Z}^{\mathcal{P}(N)}$ such that $\mathcal{B}=\mathcal{B}_\alpha$.

3.1 How to assign an inequality to a min-balanced system

Given a min-balanced system $\mathcal{B}$, unique coefficients $\lambda_S>0$, $S\in\mathcal{B}$, exist with

$$\chi_M = \sum_{S\in\mathcal{B}} \lambda_S\cdot\chi_S \qquad \text{where } M=\bigcup\mathcal{B}.$$

In fact, one can even show that $\lambda_S\in\mathbb{Q}$. Indeed, one has $\chi_M=\sum_{S\in\mathcal{B}}\lambda_S\cdot\chi_S$ means that the coefficient vector $\lambda\in\mathbb{R}^{\mathcal{B}}$ is a solution of a matrix equality $\lambda\cdot C=\chi_M$ with a zero-one matrix $C\in\mathbb{R}^{\mathcal{B}\times N}$. Since a unique solution exists, a regular column $\mathcal{B}\times T$-submatrix of C, where $T\subseteq N$, $|T|=|\mathcal{B}|$, exists such that $\lambda\cdot C^{\mathcal{B}\times T}=\chi_{M\cap T}$. Since C has zero columns for $i\in N\setminus M$ one has $T\subseteq M$. Nevertheless, the inverse of this regular zero-one submatrix is a rational matrix, which implies that the components of λ are in $\mathbb{Q}$. Thus, a unique integer $k\geq 1$ exists such that $k\cdot\lambda_S\in\mathbb{Z}$, $S\in\mathcal{B}$, are relatively prime. One can put

$$\begin{aligned}
\alpha_{\mathcal{B}}(M) &:= k,\\
\alpha_{\mathcal{B}}(S) &:= -k\cdot\lambda_S \quad \text{for } S\in\mathcal{B},\\
\alpha_{\mathcal{B}}(\emptyset) &:= -\alpha_{\mathcal{B}}(M)-\sum_{S\in\mathcal{B}}\alpha_{\mathcal{B}}(S) = -k+k\cdot\sum_{S\in\mathcal{B}}\lambda_S,\\
\alpha_{\mathcal{B}}(R) &:= 0 \quad \text{for remaining } R\subseteq N.
\end{aligned}$$

It is shown in [7, § 3.1] that these coefficients define a standardized form of the inequality (2) and one has $\mathcal{B}_\alpha = \mathcal{B}$ with $\alpha = \alpha_\mathcal{B}$. This yields mutually inverse transformation between min-balanced systems and the coefficient vectors of ascribed inequalities.

4 Irreducible min-balanced systems

The next concept is related to the conjectures below.

Definition 4 We say that a min-balanced system $\mathcal{B} \subseteq \mathcal{P}(N)$ is *reducible* if there exists $X \subset M \equiv \bigcup \mathcal{B}$ and $Y \in \mathcal{B}_X := \{S \in \mathcal{B} : S \subset X\}$ such that

- χ_X is in the conic hull of $\{\chi_S : S \in \mathcal{B}_X\}$,
- χ_M is in the conic hull of $\{\chi_T : T \in \{X\} \cup \mathcal{B} \setminus \{Y\}\}$.

A min-balanced system $\mathcal{B} \subseteq \mathcal{P}(N)$ which is not reducible is called *irreducible*.
We say that a min-balanced system $\mathcal{B} \subseteq \mathcal{P}(N)$ is *weakly irreducible* if no set $X \subset \bigcup \mathcal{B}$ exists such that both $\mathcal{B}_X$ and $\{\bigcup \mathcal{B}_X\} \cup (\mathcal{B} \setminus \mathcal{B}_X)$ are min-balanced.

Note that, without loss of generality, one can only require $X = \bigcup \mathcal{B}_X \notin \mathcal{B}$ in the above definitions and the irreducibility implies the weak irreducibility; see [7, Observation 8]. The intended meaning of the irreducibility condition is that the inequality ascribed to $\mathcal{B}$ is not derivable from other inequalities for min-balanced systems $\mathcal{B}'$ where $\bigcup \mathcal{B}' \subseteq \bigcup \mathcal{B}$.

Here is an example of a reducible system.

Example Put $N = \{a, b, c, d\}$ and consider the set system $\mathcal{B} = \{\{a\}, \{b\}, \{c\}\}$, whose corresponding inequality is

$$m(abc) - m(a) - m(b) - m(c) + 2 \cdot m(\emptyset) \geq 0\,. \tag{5}$$

Take $X = \{a, b\}$ and observe that $\mathcal{B}_X = \{\{a\}, \{b\}\}$ is min-balanced; the same holds for $\mathcal{C} = \{X\} \cup (\mathcal{B} \setminus \mathcal{B}_X) = \{\{a, b\}, \{c\}\}$. Thus, $\mathcal{B}$ is not weakly irreducible, and therefore, not irreducible. The respective inequalities are

$$\begin{aligned} m(ab) - m(a) - m(b) + m(\emptyset) &\geq 0, \\ m(abc) - m(ab) - m(c) + m(\emptyset) &\geq 0, \end{aligned} \tag{6}$$

both facet-defining for $E(N)$. Clearly, (5) is the sum of the inequalities in (6).

An analogous procedure is possible for every reducible min-balanced system. The following result is shown in [7, Corollary 9].

Observation 5 Given a reducible min-balanced system $\mathcal{B}$, the corresponding inequality is a conic combination of inequalities which correspond to other min-balanced systems $\mathcal{B}'$ with $\bigcup \mathcal{B}' \subseteq \bigcup \mathcal{B}$.

In particular, the inequalities ascribed to reducible systems are never facet-defining for $T(N)$ or $E(N)$.

5 Conjectures

The first conjecture concerns the totally balanced cone.

Conjecture 1 *The facet-defining inequalities for $T(N)$ are just those ascribed to non-trivial irreducible min-balanced systems $\mathcal{B}$ on $M \subseteq N$, $|M| \geq 2$.*

Note that $T(N)$ is not closed under reflection; therefore, one cannot expect that a conjugate inequality to a facet-defining inequality is also facet-defining.

Conjecture 2 *The facet-defining inequalities for $E(N)$ are just those ascribed to non-trivial irreducible min-balanced systems $\mathcal{B}$ on $M \subset N$, $|M| \geq 2$, and the conjugate inequalities to these.*

Conjecture 2 is in line with the fact that $E(N)$ is closed under reflection. Note that the inequalities in Conjecture 2 imply inequalities ascribed to min-balanced systems on N. Moreover, if both conjectures are true, then one can derive easily that $m \in E(N)$ iff $m, m^* \in T(N)$. That would imply that a game m is exact iff both m and m^* are totally balanced.

6 Computations and examples

A former version of the conjectures was based on weak irreducibility concept. Therefore, in case $|N| \leq 8$, we have computed the permutational types of min-balanced and weakly irreducible min-balanced systems, respectively. We listed all min-balanced systems type representatives in a tree-like catalog with the following access keys (for example, take $\mathcal{B} = \{a, bd, cd, abc\}$):

- the number $|N|$ of players $\quad (\mathcal{B}\text{: } |\{a, b, c, d\}| = 4)$,
- the number of sets $|\mathcal{B}|$ in the system $\quad (\mathcal{B}\text{: } 4)$,
- ordered players' multiplicities $|\{B \in \mathcal{B} : i \in B\}|$, $i \in N$, $\quad (\mathcal{B}\text{: } (2, 2, 2, 2))$,
- ordered cardinalities $|B|$, $B \in \mathcal{B}$, $\quad (\mathcal{B}\text{: } (1, 2, 2, 3))$,
- ordered balancing coefficients λ_S, $S \in \mathcal{B}$, $\quad (\mathcal{B}\text{: } (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}))$.

To recognize whether two given min-balanced systems are of the same type, we transformed the problem to the task of recognizing bipartite graph isomorphism. Specifically, players and sets are turned into graph nodes of two different parts. If a player is in a set, then an edge exists between the respective nodes. We have used BLISS algorithm [5], as implemented in *igraph* package [2] of R environment. To check whether a newly found min-balanced system is of a recorded permutational type (= already stored in the catalog), we searched through the leaves of the respective branch of the above tree only. Just one representative of each permutational class is stored in the catalog.

number of players	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
min-balanced types	3	9	40	428	15.309	1.597.581
weakly irreducible types	2	5	16	164	6.188	704.995

Table 1: Permutational types of non-trivial min-balanced systems on N.

To find whether the permutational type of min-balanced system is weakly irreducible, we applied Definition 4 directly. The above mentioned catalog of min-balanced systems was used to speed-up the computations.

The resulting numbers of permutation types for $3 \leq n = |N| \leq 8$ are shown in Table 1. Note that for $n \leq 4$ each weakly irreducible system is irreducible. In the case $|N| = 2$ only one non-trivial min-balanced system on N exists. It is irreducible and has the form $\mathcal{B} = \{a, b\}$ (2-partition).

In the case $n = |N| = 3$ five min-balanced systems exist which break into three permutational types; two of these types are irreducible:

- $\mathcal{B} = \{a, b, c\}$ represents a reducible type, 3-partition (1 system),
- $\mathcal{B} = \{a, bc\}$ represents an *irreducible type*, 2-partition (3 systems),
- $\mathcal{B} = \{ab, ac, bc\}$ represents an *irreducible type* (1 system).

In the case $n = |N| = 4$ one has 41 min-balanced systems which break into nine permutational types; five of these types are irreducible.

1. $\mathcal{B} = \{a, b, c, d\}$ represents a reducible type, 4-partition (1 system),
2. $\mathcal{B} = \{a, b, cd\}$ represents a reducible type, 3-partition (6 systems),
3. $\mathcal{B} = \{ab, cd\}$ represents an *irreducible type*, 2-partition (3 systems),
4. $\mathcal{B} = \{a, bcd\}$ represents an *irreducible type*, 2-partition (4 systems),
5. $\mathcal{B} = \{a, bc, bd, cd\}$ represents a reducible type (4 systems),
6. $\mathcal{B} = \{ab, acd, bcd\}$ represents an *irreducible type* (6 systems),
7. $\mathcal{B} = \{a, bd, cd, abc\}$ represents a reducible type (12 systems),
8. $\mathcal{B} = \{ab, ac, ad, bcd\}$ represents an *irreducible type* (4 systems),
9. $\mathcal{B} = \{abc, abd, acd, bcd\}$ represents an *irreducible type* (1 system).

Most of the above min-balanced systems are families of sets incomparable with respect to inclusion. Nonetheless, the min-balanced system in the 7-th item contains comparable sets a and abc.

Our computation also revealed irreducible min-balanced systems containing at least one pair of comparable sets: in case $N = \{a, b, c, d, e\}$ the system $\mathcal{B} = \{ab, acd, ace, bcde, abde\}$ is an irreducible min-balanced system.

7 Conclusions

Our future effort will be directed to the conjectures formulated in Section 5. The plan is to utilize the methods of polyhedral geometry either to confirm or to disprove them. The catalogues of irreducible min-balanced systems for $|N| = 6, 7, 8$ are highly useful in this context, because they determine the lists of inequalities conjectured to be facet-defining for $T(N)$ and $E(N)$.

In another direction of research related to the cones $\mathcal{S}(N)$ and $\mathcal{E}(N)$, we found simple linear criteria to recognize extreme supermodular/exact games and developed web platforms based on implementation of these criteria [14, 15].

Acknowledgement

This work was supported by the Czech Science Foundation (project 16-12010S).

References

[1] O. Bondareva: Some applications of linear programming methods to the theory of cooperative games (in Russian). Problemy Kibernetiki 10 (1963) 119-139.

[2] G. Csardi, T. Nepusz: The igraph software package for complex network research, InterJournal Complex Systems 1695, 2006. Available through `http://igraph.org`.

[3] P. Csóka, P. J.-J. Herings, L. A. Kóczy: Balancedness condition for exact games. Mathematical Methods of Optimization Research 74 (2011) 44-52.

[4] M. Grabisch: Set Functions, Games and Capacities in Decision Making. Springer 2016.

[5] T. Junttila, P. Kaski: Engineering an efficient canonical labeling tool for large and sparse graphs. Proceedings of the 9-th on Algorithm Engineering and Experiments and the 4-th Workshop on Analytic Algorithms and Combinatorics, 2007, 135-149.

[6] J. Kuipers, D. Vermeulen, M. Voorneveld: A generalization of the Shapley-Ichiishi result. International Journal of Game Theory 39 (2010) 585-602.

[7] J. Kroupa, M. Studený: Towards facets of the exact cone. A manuscript in preparation (2018).

[8] E. Lohmann, P. Borm, P. J.-J. Herings: Minimal exact balancedness. Mathematical Social Sciences 64 (2012) 127-135.

[9] E. Miranda, I. Montes: Games solutions, probability transformations and the core. In JMLR Workshops and Conference Proceedings 62: ISIPTA 2017, 217-228.

[10] B. Peleg: An inductive method for constructing minimal balanced collections of finite sets. Naval Research Logistics Quaterly 12 (1965) 155-162.

[11] E. Quaeghebeur: Learning from samples using coherent lower probabilities. PhD thesis, Ghent University 2009. Available through http://hdl.handle.net/1854/LU-495650 .

[12] L. S. Shapley: On balanced sets and cores. Naval Research Logistics Quarterly 14 (1967) 453-460.

[13] M. Studený: Probabilistic Conditional Independence Structures. Springer 2005.

[14] M. Studený, T. Kroupa: Core-based criterion for extreme supermodular functions. Discrete Applied Mathematics 20 (2016) 122-151.

[15] M. Studený, V. Kratochvíl: Linear core-based criterion for testing extreme exact games. In JMLR Workshops and Conference Proceedings 62: ISIPTA 2017, 313-324.

Employing Bayesian Networks for Subjective Well-being Prediction

Jan Švorc[1,2] and **Jiří Vomlel**[1]
xsvoj16@fm.vse.cz, vomlel@utia.cas.cz

[1]Institute of Information Theory and Automation
Czech Academy of Sciences

[2]Faculty of Management
University of Economics, Prague

Abstract

This contribution aims at using Bayesian networks for modelling the relations between the individual subjective well-being (SWB) and the individual material situation. The material situation is approximated by subjective measures (perceived economic strain, subjective evaluation of the income relative to most people in the country and to own past) and objective measures (household's income, material deprivation, financial problems and housing defects). The suggested Bayesian network represents the relations among SWB and the variables approximating the material situation. The structure is established based on the expertise gained from literature, whereas the parameters are learnt based on empirical data from 3rd edition of European Quality of Life Study for the Czech Republic, Hungary, Poland and Slovakia conducted in 2011. Prediction accuracy of SWB is tested and compared with two benchmark models whose structures are learnt using Gobnilp software and a greedy algorithm built in Hugin software. SWB prediction accuracy of the expert model is 66,83%, which is significantly different from no information rate of 55,16%. It is slightly lower than the two machine learnt benchmark models.

1 Introduction

Throughout the last couple of decades, subjective well-being (hereinafter referred as "W") has become an attractive field of study for sociologists, psychologists and economists. Each of these branches looks at the topic from the different perspective, whereas their models reflects mainly their own needs and understanding. For example, economists use the concept of SWB in their models as a certain approximation when investigating the utility. The empirical research of SWB and its

association with economic variables such as individual income, material situation, relative deprivation, material deprivation etc. is usually based on the use of various statistical methods. Applications of probability calculus and modelling to examine sociological concepts such as SWB seem to be rather marginal. In this contribution we tried to develop an alternative to the classical statistical approach and suggest the probability model for the description of the relationships among the individual SWB and selected proxies of the economic situation based on empirical data from four central European countries. For such an attempt Bayesian networks are used, whereas the predictive capacity of the model is discussed in terms of SWB. The broad purpose of this study is to demonstrate possible new approaches for modelling in sociology.

2 Literature

SWB can be defined uneasily and approaches to that differ quite a lot. For the purposes of this research the definition of SWB promoted in [9] is used. Based on this definition there are two components of SWB, affective and cognitive. The cognitive component can be understood as a judgement of one's life satisfaction, whereas the affective dimension is represented by emotions and moods. The surplus of positive emotions over negative ones is referred as happiness. Comprehensive overviews of the issues defining SWB are provided for example in [7] or [10]. Both outlined dimensions of SWB are reflected in the empirical analysis in this paper.

The individual material situation can be approximated by diverse variables, both objective and subjective. The term "material situation" appears in the study [4] where direct and indirect measures are discussed. Economic variables once examined were mostly limited to the income [22]. Household's income is the most obvious indirect measure, but the evidence of the relation between the income and SWB is mixed. Some authors stress the importance of the income for SWB whereas others look it rather unimportant. The reviews of the evidence are provided eg. in [7], [9] and [25]. Overall, researchers mostly suggest that money has a positive, yet diminishing effect on SWB [11, p. 97], which is in line with the widely accepted economic law of declining marginal utility.

Some studies (e.g. [9]) suggest that the relative, rather than the absolute, income matters, as people simply tend to compare one to another. Clark et al. [5] talk about the comparison to others and the comparison to oneself in the past. Diener et al. [8, p.195] summarize that the impact of the income depends on "changeable standards derived from expectancies, habituation levels, and social comparisons". It means that the additional income has no effect on SWB if the income of people in the reference group also increase [11, p. 98]. With some exemptions (e.g. [8]), the research is quite consistent in the assertion of importance not to consider the effect of income to SWB only in absolute terms.

Material deprivation measures are used in poverty research since introduced in [3] and [30]. It is commonly assumed that there is a close relationship between

the income and the material deprivation as the lack of resources caused by the low income results in the lack of something considered to be a necessity. However, the research often suggest that the discrepancies between the income and the material deprivation exist. Studies suggesting that there are low income households not experiencing deprivation as well as households not living in poverty but suffering from the deprivation (measured by non-monetary indicators) are summarized, for example, in [32]. The authors provided reasoning why the income and the material deprivation may relate loosely – the length of time the low income persists, the existence of other resources (savings etc.), a different view of what is necessary, and other social and economic processes may influence the relationship between current income and deprivation. They made analysis of the income, the deprivation and the economic strain based on data for twelve European Union countries from the first wave of the European Community Household Panel Study (ECHP) conducted in 1994 and found that the relationship of the income with the deprivation "was generally weakest in the richer countries where the level of deprivation is lowest, and strongest where it is highest" [32, p. 370] and the economic strain (the perceived ability to make ends meet) is impacted by both the income and the deprivation whereas the effect of the deprivation is stronger. The structure of the deprivation is consistent across examined EU countries. Based on European Quality of Life Survey (EQLS) data for EU25 plus 3 candidate states at that time (Romania, Bulgaria and Turkey) [33] later confirmed the earlier finding that the relationship between the income and the life style deprivation is relatively weak, whereas the income plays more important role in the poorer regions as a predictor of deprivation. The association between the deprivation and the economic strain is the strongest in 12 richest EU countries whereas in all other EU regions is just a bit weaker.

3 Data and method

Learning the structures and the parameters is based on the empirical data from European Quality of Life Study (hereinafter referred as "EQLS") carried out by the European Foundation for the Improvement of Living and Working Conditions – Eurofound [19] covering all 27 EU member states and 7 non-EU countries. Third edition of the survey conducted in Autumn/Winter 2011 is used for this purpose. Only one interview per household was held whereas the adult household member with the next upcoming birthday was taken as the eligible respondent. The statistical population of the study covered all persons aged 18 and over whose usual place of residence was in the territory of the surveyed country. Random probability sampling procedures were used promising that every member of the statistical population have non-zero probability to be included in the sample. The sample was stratified according to NUTS2, level of urbanization and clustered geographically on Primary Sampling Units. The sample can hence be considered representative. All necessary technical details how the study was conducted are available in EQLS Technical report [18] and EQLS Sampling report [17].

Probability models presented in this contribution are learnt using only data of four post-communist central European countries (the Czech Republic, Hungary, Poland and Slovakia). This approach promises to have a sufficient sample from very similar countries in terms of culture, geography, politics, economics and modern history. Total sample size for the four countries is 5.298 (1,012 in the Czech Republic, 1.024 in Hungary, 2.262 in Poland and 1.000 in Slovakia) respondents out of whom 3.797 (722 in the Czech Republic, 687 in Hungary, 1.707 in Poland and 681in Slovakia) complete data vectors are extracted by removing respondents not having answered the relevant questions. This selection of countries is represented by the node COUNTRY in the model.

Broad range of domains is covered by EQLS including SWB and the financial situation measures. There is no single question on SWB, there are questions on both happiness and life satisfaction instead. For the happiness respondents are asked in the following way: "Taking all things together on a scale of 1 to 10, how happy would you say you are?" Code 1 means very unhappy and 10 means very happy on the scale. The question of the overall satisfaction is: "All things considered, how satisfied would you say you are with your life these days?" Similarly to happiness the scale of 1 to 10 is given, where 1 means very dissatisfied and 10 means very satisfied. Model variable of the subjective well-being (abbreviated as "AVGSWB" in the model) is binary where one state represents SWB below the median and the other state represents SWB on or above the median. For each respondent the average of happiness and life satisfaction is computed first and then the median is calculated from this working scale. Because SWB is a two-item measure, the internal consistency was checked using Cronbach's alpha. The value of alpha is 0.774 which is generally considered acceptable in the social research.

A set of seven variables is used for the description of the individual material situation and the deprivation in the model. Household income, material deprivation, defects in housing conditions and financial problems are the objective ones. The subjective variables comprise the subjective economic strain and the relative income approximated by the subjective evaluation of own current financial situation compared to other people in the country and by the subjective evaluation of own current financial situation compared to own past situation. People are often cautious stating their household income (total of 1.333 respondents out of the 5.298 did not declared their income). The respondents were asked to state their total net income per month from all sources of all members of the household. The household income equalized based on purchasing power parity euros is used for the analysis, such a figure is provided in the EQLS dataset. The variable of income is binary where one state covers the income up to the median (including) and the second state the income above median.

EQLS respondents were asked whether they are able to afford six items if they would like them to get. The six items include keeping your home adequately warm, paying for a week's annual holiday away from home (not staying with relatives), replacing any worn-out furniture, having a meal with meat, chicken, fish every second day (if wanted), buying new, rather than second-hand, clothes and having

friends or family for a drink or meal at least once a month. Resulting six binary variables (able to afford / unable to afford) are transformed into a single binary variable, where one group of respondents can afford all six items and the other group cannot afford one to six items. The respondents having refused to answer the whole set of six sub-questions are excluded from the further analysis. This variable of the material deprivation is abbreviated as "WANTED" in the model.

Financial problems in the form of the ability to pay various households bills as scheduled in the past 12 months is another objective variable related to the living conditions. The question on the financial problems has five sub-questions covering payments that typical household needs to pay regularly: rent or mortgage payments for accommodation, utility bills, such as electricity, water, gas, payments related to consumer loans, including credit card overdrafts and payments related to informal loans from friends or relatives not living in the household. Four binary variable (able to pay bills as scheduled / unable to pay bills as scheduled) are transformed into a single binary variable where one state means able to afford all four items whereas the other state means unable to afford one to six items. Consistent principle is adopted for those not answering questions. Excluded are those who did not answer all the sub-questions. This variable of the financial problems is abbreviated as "UNABLEPAY" in the model. The last objective variable related to the material living conditions describes defects in housing conditions. The related EQLS question has six sub-questions on type of common deficiencies of housing: shortage of space, rot in windows, doors or floors, damp or leaks in walls or roof, lack of indoor flushing toilet, lack of bath or shower and lack of place to sit outside (e.g. garden, balcony, terrace etc.). Single binary variable is again derived from the six partial binaries (problem exists / problem does not exist) same way (one category contains only respondents having no problems with housing and the other contains respondents having one to six problems with housing, respondents having refused all six items are excluded from the dataset). This variable of the housing defects is abbreviated as "ACCOMP" in the model.

Subjective measures are based on the respondent's subjective feeling of their own situation rather than on the objective material living conditions. Broadly speaking the relative income is the income compared with a defined standard given by other incomes as perceived by the respondent. As discussed before the income can be compared with own income in the past and the income of people in the country, region or closer neighborhood. Same for the financial situation. The EQLS question dealing with own financial situation compared with others is the following: "Could you please evaluate the financial situation of your household? In comparison to most people in your country would you say it is much worse, somewhat worse, neither worse nor better, somewhat better or much better?" For the purpose of modelling this 5-point scale is transformed into the 3-point scale (worse, the same and better financial situation). The variable is hence ternary and is abbreviated as "FINSITEVAL". The EQLS question on comparison with own past was posed this way: "When you compare the financial situation of your household 12 months ago and now, would you say it has become better, worse or

remained the same?" The ternary variable is abbreviated as "PASTFIN" in the model.

The perceived economic strain is covered by the following EQLS question: "Thinking of your household's total monthly income: is your household able to make ends meet very easily, easily, fairly easily, with some difficulty, with difficulty or with great difficulty?" This 6-point scale is transformed to only two categories of those able to make ends meet easily and those able to make ends meet with difficulty. The transformed variable used for modelling is abbreviated as "MEETENDS".

The model in the form of Bayesian network is constructed using the above outlined data. The expertise gained from the existing literature review is used to establish the structure of the model, whereas parameters are learnt using EQLS data as described above. To evaluate the predictive accuracy of the expert model two more benchmark models are constructed using different structure learning approaches: a greedy algorithm and the optimal Gobnilp algorithm [6]. To summarize, in this empirical study, following models are considered:

- **Greedy-BIC** – structure learnt with the greedy search-and-score algorithm with the BIC scoring criterion.
- **Gobnilp-BIC** – structure learnt using the Gobnilp algorithm with the BIC scoring criterion.
- **Expert** -– structure learnt using expert knowledge.

he structural learning of the Greedy-BIC model is performed in the analytic software Hugin [23], parameters of all models are learnt in Hugin as well. Structure of the Gobnilp-BIC model is learnt using Gobnilp software [6]. Gobnilp-BIC model is optimal in terms of the BIC criterion.

4 Model and discussion

In Figure 1 we present the suggested Bayesian network which represents the relations among SWB and the variables approximating the material situation of an individual. The variable referred as COUNTRY represents the country the respondent come from. Figure 1 Bayesian network structure (expert version). The suggested model is examined from the perspective of conditional independencies and the ability to predict SWB based on the given material situation variables. Expert argumentation for the relations between the nodes is summarized with the special respect to SWB.

4.1 4.1. Variables directly linked to SWB

SWB is directly linked with the relative income expressed as both *the income relative to own past* (represented by the node PASTFIN) and *the income relative*

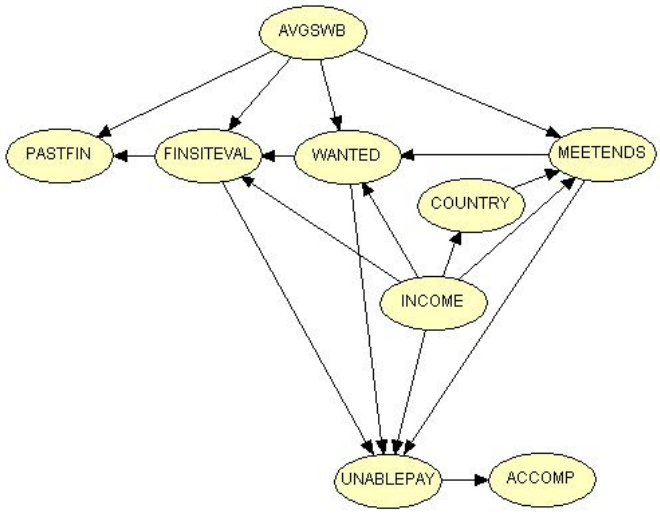


Figure 1: Bayesian network structure (expert version)

to the other people in the country (represented by the node FINSITEVAL). The two expressions of the relative income are also directly linked. As seen before, the direct link between SWB and the relative income can be traced to the literature, for example [5] and [8]. Simply said, people are unhappy and unsatisfied when feeling their material situation get worse comparing to either what was before or what others have. The measures are both subjective. SWB is also directly linked with the *material deprivation* (represented by the node WANTED). This direct link can also be supported by the literature, for example [1]. Basically SWB drops down if one hunger after something that cannot be afforded. The last variable directly linked with SWB is *perceived economic strain* (represented by the node MEETENDS). The evidence for the direct link can again be found in the literature, for example [26] and [12]. Common sense reasoning is that SWB is reduced in case a household is unable to make ends meet. The direct link between the *economic strain* and the *material deprivation* can be supported by the literature too, where [32] and [33] stand as examples. Inability to make ends meet in essence corresponds directly with the impossibility to afford things desired. Similarly, the subjective assessment of own financial situation compared to most people in the country is linked with the *material deprivation* – people are deprived as they cannot afford things they think other people in the country mostly can. To summarize, SWB is directly linked with all other subjective measures in the model (there is no other subjective measure in the model). The only objective measure directly linked with SWB is the *material deprivation*. These five interconnected variables constitutes a sort of cluster within the network.

4.2 WB and income and country

Two variables are d-separated from SWB in the model – the country and the income. It means that if we know nothing else, SWB is independent of the country and the income. It can be easily argued that the household's income in absolute terms is linked with the country in which the household resides. It is not that apparent from the perspective of the four examined countries, but it is clear globally. Independence of SWB and the country means the level of SWB cannot simply imply the characteristics of a given country and vice versa. For example, we cannot conclude that people living in the rich countries are automatically happier and more satisfied with their lives than those living in the poor countries although the common sense might suggest otherwise.

Modern economic discussion on the relations between the level of SWB and the economic performance of a country as well as household's level of incomes was launched by Richard Easterlin [13]. Based on the empirical evidence he proposed, that there is a noticeable positive association of the income with the happiness within a given country. But the picture is quite different from the international perspective; the reported level of happiness on average was not associated with national income per head. At least in countries, where the income per head is sufficient to cover basic needs[1]. Shortly, people in a rich country are not on average happier than people in a poor country (given that basic needs are met), but within a given country the income and the reported happiness are associated. Furthermore, according to the Easterlin, there is no correlation between the increase in national product per person and the increase in the average reported happiness over the long-term. Twenty years later Easterlin [14] reacted to the criticism of his research (e.g. [31]) and refined his original conclusion. This refinement is that raising the incomes of all people in the society does not increase the happiness of all, because, so-called, "material norms" increase in the same proportion as the income of the society. When the country becomes richer, higher level of income becomes normal – although the absolute income is higher, the relative income remains the same. This Easterlin's arguments are in essence in line with the conclusions of the importance of the relative income explored above. Once again Easterlin [16] confirmed his paradox on updated dataset and stressed that the happiness and the economic performance are not related only in long term, while short term fluctuations of the happiness and the national income are positively associated. Because of confusion the short and long term trends, some authors may suggest the positive relation of SWB and GDP in long term too. There are many studies confirming (e.g. [2]) and disputing Easterlin's conclusion (e.g. [21], [31]) and the debate is certainly not over.

[1] His conclusion was later nicknamed Easterlin paradox.

4.3 Variables intermediating SWB and income

Based on the suggested graph SWB and the country are d-connected when other information is available; information of either relative income or material deprivation or perceived economic strain. The importance of the relative income as a mediator between SWB and the economic performance of the country has just been discussed. If the relative income is known, the conclusion on the country can theoretically be drawn from the level of SWB and vice versa. Still, this conclusion can be reached rather in case of more different countries than the four examined central European ones[2]. The perceived economic strain as well as the material deprivation could be contemplated in the same way as mediators. Both variables were examined in the EU-wide context by [20], who suggest that these measures should be employed as indicators for certain purposes rather than the income as they provide better information that pure income thresholds in situation of inequality in the income between the EU member states.

The important feature of the model is that the absolute level of the income is treated as conditionally independent of SWB and the knowledge of mediating factors is needed since SWB and the income are d-separated (given an empty set). It is a sort of the model assumption, as immense amount of studies confirming that the income and SWB of an individual are associated can be found as already discussed. Still, newer research tend to understand the income rather as indirect measure (e.g. [4]). Relative income, perceived economic strain and material deprivation are the suggested mediators between the income and SWB in the Bayesian network. If we know one of them, the link from the income to SWB is unblocked. Direct links from the income to all of the three mediating variables can be well argued.

In case of higher current level of the income the probability of the subjective evaluation of own income compared to others increases. (Absolute income is not directly linked with the relative income compared to past, the way goes through the node FINISTEVAL. It means the knowledge of the income is not necessary to conclude on the relative income compared to past). Similarly higher income can be directly associated with lower perceived economic strain (making ends meet is more probable) as well as with less material deprivation (affordability of things wanted is more probable). Same the other direction. It is assumed that the income and SWB is independent, but if we know how this income is perceived relative to other people in the community, we can conclude on SWB and vice versa. Similarly, with the knowledge of the material deprivation or the perceived economic strain we can conclude on SWB from the income. For example high income leading to the ability to afford things wanted leads us to conclude on good level of SWB. On the other hand the same level of the income can be insufficient to afford things wanted for another household which will lead to dissatisfaction as they become deprived. We must have such information to conclude on SWB from the income.

[2]For example, such a conclusion can be made in case of comparison the Czech Republic with Bhutan, a poor Buddhists country known for their extraordinary nation-wide approach of pursuing happiness.

This is confirmed by the literature concerning on material aspirations. Based on [15] the material aspirations increase together with the income over the life course, whereas SWB, generally, rises with the income, but inversely with the material aspirations. The rise of the income causes the rise of SWB on the one hand, but also the rise of the material aspirations on the other hand. The rise of the material aspirations affects negatively SWB and erase the positive effect of the income. People tend to want more and more throughout the life, which negatively affects their SWB (SWB would otherwise gain from the increase of the income). Similarly [29] empirically tested the effect of the income aspirations on people's utility operationalized as reported satisfaction with life. He founds that, ceteris paribus, higher income aspirations reduce people's utility measured by the satisfaction with life. The author offers two explanations for that: processes of adaptation and social comparison. Firstly, the increase of the income initially provide additional pleasure at the beginning, but the effect disappear as people get used to the new income level. Secondly, the relative income position rather than absolute level of the income matters, because people tend to compare themselves with others in the community. Stutzer talks about "socially comparative or even competitive processes in consumption" [29, p.3]. This view is in line with previously referred Easterlin paradox, because people in a rich country are adapted to their material standards and simple fact that they are richer than people in a poor country cannot make them happier. People in the poor country have their own standards they are adapted to. To conclude on the material aspirations, [28] found the negative relationship of life-satisfaction with materialism (which could be seen as individual orientation to possession and acquisition). Materialistically oriented people were less satisfied with their lives as a whole, with their standard of living, family lives and other life domains than those low in materialism.

The literature hence confirmed that it is important to know whether the income is sufficient for the ability of afford desired thing in order to conclude on SWB. Logic is the similar in case of the perceived economic strain. The income sufficient to safely make ends meet probably promotes the higher level of SWB, still, the other household may not be able to manage to make ends meet with the income in the same amount.

4.4 Other variables

Only variables examined so far and no other variables are in the Markov blanket[3] of the node SWB, the two of them (income and country) are d-separated given the empty set and the rest is d-connected (given the empty set). If we know the states of them, no other information is necessary to conclude on SWB. The two other variables remains outside the Markov blanket: *financial problems* (UNABLEPAY) and *housing defects* (ACCOMP). They are both associated with SWB, but having

[3]The Markov blanket of a variable A is the set consisting of the parents of A, the children of A, and the variables sharing a child with A. If all variables in the Markov blanket for A are instantiated, then A is d-separated from the rest of the network [24, p. 11].

the information of either relative income or perceived economic strain or material deprivation, information of the financial problems (and the housing defects) are not necessary. If a household does not feel economically strained (or feels their situation is relatively good or can afford what they otherwise want), the SWB of that household might be good even if it has problems paying bills in arrears. On the contrary, if there are no problems paying bills, but members of the household feel deprived, economically strained and bad in terms of their situation compared to others, their SWB will probably be worsen. Still, if we have no other knowledge, information on financial problems is relevant in terms of SWB prediction.

Other links in the model can be explained in similar way and analyzed further. For example the income is directly linked with financial problems – having less income may affect the ability to pay bills if households has no savings. The last comment belongs to the node of housing defects, which seems to stay a bit apart from the heart of the network as it is linked with the rest of the network only via the node of financial problems. This variable covers, inter alia, the problems with payments of rents, mortgages, utility bills etc. It could hence be expected that the inability to pay such bills is also connected with housing defects (household has to move to a smaller apartment of a lower standard).

4.5 Prediction of SWB

We perform an evaluation of the prediction accuracy of the expert model and compare it with two learning algorithm approaches. The prediction accuracy is tested using the R software [27]. The adopted approach uses 75% of the available dataset as training data to learn the parameters of the model whereas the obtained predictions of SWB are compared with the true state of SWB on remaining 25% observations in the second step. Ten rounds of such tests were performed. The average accuracy of the expert model in predicting SWB is 66.83% (95% confidence interval 65,87% to 67,78%), which is significantly different from no information rate of 55.16%.

The prediction accuracy of the expert model is compared with the other two models presented in Figures 2 and 3. The SWB prediction accuracy of the Greedy-BIC model learnt with Hugin [23] is 67.01% (the 95% confidence interval is from 66.05% to 67.95%). In case of the Gobnilp-BIC model learnt with GOBNILP software [6] the prediction accuracy of SWB is 67.00% (the 95% confidence interval is from 66.04% to 67.94%). This third model is optimal with respect to the Bayesian Information Criterion (BIC) and it has its BIC equal to -24390.4 while Greedy-BIC model reached BIC of -24391.2. The expert model is worst out of the three suggested model with BIC equal to -24754.2. The two machine learnt models are almost identical in terms of the predictive accuracy as well as BIC, where the Gobnilp-BIC model seems to be only marginally better as suggested by both measures.

Although expert model is the weakest in terms of BIC and the predictive accuracy, it does not significantly differ from the other two models learnt from data.

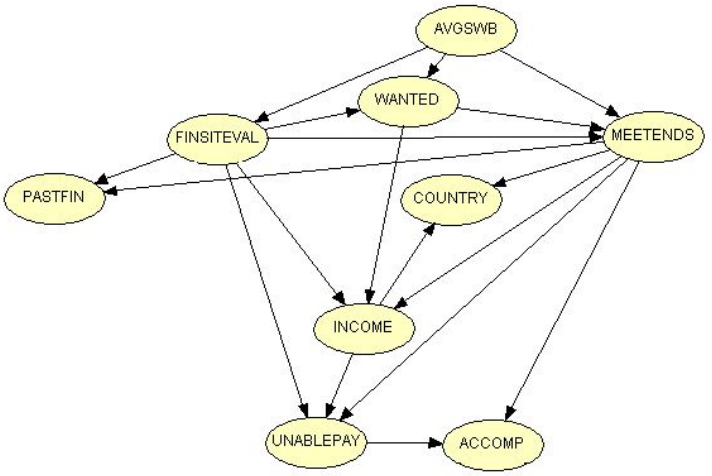


Figure 2: Bayesian network structure (Greedy-BIC version)

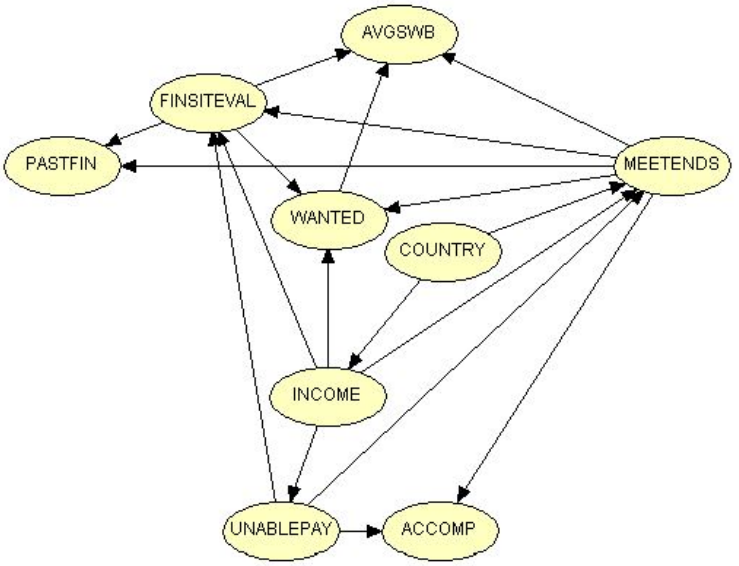


Figure 3: Bayesian network structure (Gobnilp-BIC version)

The expert version is worth considering, because it represents relations and conditional independencies of SWB and factors related to material conditions known from the social research made so far probably better that the other two models. The key difference between the expert version and the other two models is the conditional independence of the income and SWB as well as the country and SWB. The key feature of the expert version is that the income (the country) and SWB are conditionally independent (given the empty set) as explained earlier, whereas the income and SWB as well as country and SWB are d-connected in the benchmark models.

5 Concluding remarks

This contribution is an attempt to employ Bayesian networks in a research of sociological topics such as SWB. To the knowledge of the authors the Bayesian network approach has not yet been adopted in the way described in this paper. However, as such, it should be understood as a first step of a longer journey.

For the analysis, the EQLS data of four Central European countries collected in 2011 are used with the argument of the similarity of these countries from several points of view. The model hence reflects the situation in these post-communist countries and a care should be taken when using it in different context. Recently, newer data were made available[4], there is hence a room for updating and a further exploration.

Expert model, as well as the other two models, are able to predict SWB based on material living conditions and deprivation considering that these factors constitute only a small part of the whole picture. Immense number of studies is available on SWB and how it is associated with the factors related to demography, aspirations, expectations, personality, social relations and wider environment, where personality traits seems to matter in long term, while life events play the role rather in short term. Most of the factors we examined falls into the latter category, whereas the subjective opinion of material living standards and deprivation are certainly impacted by the personality too. Other than material factors are not reflected in the suggested network as they are too tangled to be described in their complexity.

Acknowledgement

This work was supported by the Czech Science Foundation (project 17-08182S).

References

[1] L. Bellani and C. D'Ambrosio. Deprivation, social exclusion and subjective well-being. Social indicators research, 104(1): 67-86, 2011.

[4] 4th edition of EQLS conducted in 2016 and released in 2018.

[2] D. G. Blanchflower and A. J. Oswald. Well-being over time in Britain and the USA. Journal of public economics, 88(7): 1359-1386, 2004.

[3] V. Carstairs and R. Morris. Deprivation: explaining differences in mortality between Scotland and England and Wales. British Medical Journal, 299(6704): 886-889, 1989.

[4] B. Christoph. The relation between life satisfaction and the material situation: A re-evaluation using alternative measures. Social Indicators Research, 98(3): 475-499, 2010.

[5] A.E. Clark, P. Frijters and M.A. Shields. Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles. Journal of Economic Literature, 46(1): 95-144, 2008.

[6] J. Cussens and M. Bartlett. GOBNILP software. 2015.

[7] E. Diener. Subjective Well-Being. Psychological Bulletin, 95(3): 542-575, 1984.

[8] E. Diener, E. Sandvik, L. Seidlitz and M. Diener. The relationship between income and subjective well-being: Relative or absolute? Social Indicators Research, 28(3): 195-223, 1993.

[9] E. Diener, E. Suh, R. E. Lucas and H. L. Smith. Subjective well-being: Three decades of progress. Psychological Bulletin, 125(2): 276–302, 1999.

[10] R. Dodge, A. P. Daly, J. Huyton and L. D. Sanders. The challenge of defining wellbeing. International Journal of Wellbeing, 2(3): 222-235, 2012.

[11] P. Dolan, T. Peasgood and M. White. Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. Journal of economic psychology, 29(1): 94-122, 2008.

[12] H. Ervasti and T. Venetoklis. Unemployment and subjective well-being: An empirical test of deprivation theory, incentive paradigm and financial strain approach. Acta Sociologica, 53(2): 119-139, 2010.

[13] R. A. Easterlin. Does economic growth improve the human lot? Some empirical evidence. Nations and households in economic growth, 89: 89-125, 1974.

[14] R. A. Easterlin. Will raising the incomes of all increase the happiness of all? Journal of Economic Behavior & Organization, 27(1): 35-47, 1995.

[15] R. A. Easterlin. Income and happiness: Towards a unified theory. Economic journal, 111(473): 465-484, 2001.

[16] R. A. Easterlin. Happiness and economic growth–the evidence. Global handbook of quality of life: 283-299, 2015.

[17] European Foundation for the Improvement of Living and Working Conditions, Third European Quality of Life Survey: sampling report, 2012.

[18] European Foundation for the Improvement of Living and Working Conditions, Third European Quality of Life Survey: technical report, 2012.

[19] European Foundation for the Improvement of Living and Working Conditions, European Quality of Life Survey [computer file], 2011-2012: 2nd Edition. Colchester, Essex: UK Data Archive [distributor], January 2014. SN: 7316, http://dx.doi.org/10.5255/UKDA-SN-7316-2.

[20] T. Fahey. The Case for an EU-wide Measure of Poverty. European Sociological Review, 23(1): 35-47, 2006.

[21] M. R. Hagerty and R. Veenhoven. Wealth and Happiness Revisited: Growing National Income Does Go with Greater Happiness. Social Indicators Research, 64(1): 1-27, 2003.

[22] B. Headey and M. Wooden. The Effects of Wealth and Income on Subjective Well-Being and Ill-Being. Institute for the Study of Labor (IZA), Bonn, IZA Discussion paper series, 1032, 2004.

[23] Hugin Expert, Hugin Researcher API 8.0 (www. hugin. com), 2014.

[24] F. V. Jensen. Bayesian networks and decision graphs. Aalborg University, Aalborg, 2001.

[25] R. Larson. Thirty years of research on the subjective well-being of older Americans. Journal of gerontology, 33(1): 109-125, 1978.

[26] R. J. Mills, H. G. Grasmick, C. S. Morgan and D. Wenk. The effects of gender, family satisfaction, and economic strain on psychological well-being. Family Relations, 41(4): 440-445, 1992.

[27] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2014.

[28] L. Ryan, S. Dziurawiec. Materialism and its relationship to life satisfaction. Social Indicators Research, 55(2): 185-197, 2001.

[29] A. Stutzer. The role of income aspirations in individual happiness. Journal of Economic Behavior & Organization, 54(1): 89-109, 2004.

[30] P. Townsend, P. Phillimore and A. Beattie. Inequalities in health in the Northern region. Newcastle upon Tyne: Northern Regional Health Authority and University of Bristol, 1986.

[31] R. Veenhoven. Is happiness relative? Social indicators research, 24(1): 1-34, 1991.

[32] C. T. Whelan, R. Layte, B. Maître and B. Nolan. Income, deprivation, and economic strain. An analysis of the European community household panel. European sociological review, 17(4): 357-372, 2001.

Dynamic Bayesian Networks for the Classification of Sleep Stages

Jiří Vomlel[1,2] and **Václav Kratochvíl**[1,2]
vomlel@utia.cas.cz, velorex@utia.cas.cz

[1]Institute of Information Theory and Automation
Czech Academy of Sciences

[2]Faculty of Management
University of Economics, Prague

Abstract

Human sleep is traditionally classified into five (or six) stages. The manual classification is time consuming since it requires knowledge of an extensive set of rules from manuals and experienced experts. Therefore automatic classification methods appear useful for this task. In this paper we extend the approach based on Hidden Markov Models by relating certain features not only to the current time slice but also to the previous one. Dynamic Bayesian Networks that results from this generalization are thus capable of modeling features related to state transitions. Experiments on real data revealed that in this way we are able to increase the prediction accuracy.

1 Introduction

Human sleep occurs in cycles. Each cycle lasts approximately 90 minutes. Typically, there are four or five cycles per night. Traditionally, sleep stages are scored into the following five stages of sleep: (W) Wakefulness, (REM) the stage named after "rapid eye movement" that occurs during this stage, and three Non-REM stages named (N1), (N2), and (N3). In Figure 1 we present an example of a real hypnogram from the dataset we used in our experiments.

Traditionally, a set of rules from a manual is used for sleep scoring. The rules for sleep scoring were standardized by [5]. An up-to-date manual is provided by the American Academy of Sleep Medicine, see [3]. However, it should be beneficent if the scoring system is based on probabilistic principles rather than on a predefined set of rules. Hidden Markov models (HMMs) can play this role.

In [4] the authors verify on synthetic data that Gaussian Observation HMMs (GOHMMs) can detect the state transitions and are thus a model well-suited for

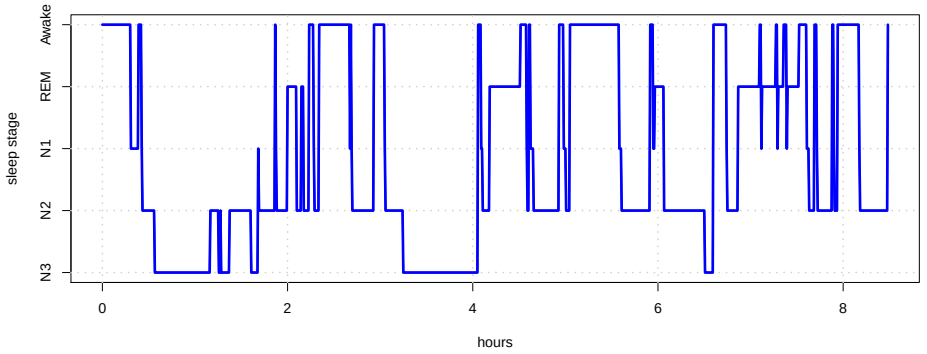


Figure 1: An example of a hypnogram.

the EEG analysis.In [2] a probabilistic continuous sleep stager based on Hidden Markov models is developed using only a single EEG signal. In the work of [1] an automatic diagnosis system based on HMMs is proposed to help clinicians in the diagnosis of sleep apnea syndrome.

In our paper we build on these works and use a generalization of HMMs – Dynamic Bayesian networks. This allows us to properly include features related to hidden states of two consecutive stages into the model. We used one such feature from our dataset – Arousal, which appears typically when a transition between sleep stages occurs. We will see that already treating a single feature properly (in a DBN model) helps us to significantly increase the prediction accuracy.

In this paper, first, we discuss probabilistic models suitable for the sleep classification task. We start with two versions of the Naive Bayes (NB) model in Section 2. In Section 3 we extend the NB model in two steps. First, the time factor is included into the model by considering the transition probabilities between sleep stages – in this way we get Hidden Markov Models (HMMs). Second, the HMMs are generalized by relaxing the Markov property. HMMs are generalized to Dynamic Bayesian Networks (DBNs) by including features related to transitions between two consecutive stages. We conclude the paper by experimental evaluation of the considered probabilistic models (in Section 4), by a summarizing our results, and by a discussion of possible future work (in Section 5).

2 Naive Bayes models

Probabilistic models offer an advantage of being objective by not relying on human scorers and being based on solid probabilistic principles rather than a predefined set of rules. A Naive Bayes (NB) model is an example of a simple probabilistic

model. It is a Bayesian network model that builds on the assumption of conditional independence of observed features given the state (class) variable. In the context of sleep analysis, it means that the observed attributes are assumed to be independent given the sleep stage. The model is static in the sense that features observed in time t are relevant for the classification of the sleep at time t only.

In Figure 2 we present an example of a Naive Bayes model. Variable Y represents the sleep stage at a given time t and variables $X_1, \ldots, X_5$ are features. Examples of features used in the sleep analysis are:

- the spectral density at given frequency ranges of the Electro Encephalogram (EEG) signal,
- Electro Oculogram (EOG), which identifies eye movements,
- Electro Myogram (EMG), which identifies muscle activity,
- Central Sleep Apnea,
- Snoring, etc.

Most of these features are continuous variables (in figures displayed as double circles) and some of them are discrete – often binary (in figures displayed as single circles).

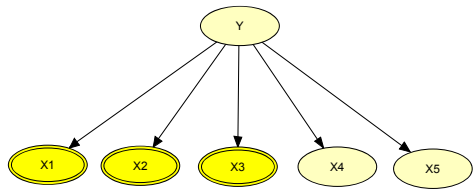


Figure 2: The structure of the Naive Bayes Model for clock time t.

For continuous features X a natural model for the conditional density $P(X = x|Y = y)$ seems to be the Conditional Gaussian distribution $\mathcal{N}(\mu_y, \sigma_y)$ where μ_y and σ_y are the mean and the standard deviation of feature X given the value y of parent variable Y. At the top of Figure 3 we can see the Gaussian density estimates for the delta power spectral density learned from real data.

In the experiments (reported in Section 4) it appeared that the CGDs did not lead to a good prediction accuracy. A better option appeared to be the Kernel Density Functions (KDFs) that can better fit the actual shape of the conditional density functions. At the bottom of Figure 3 we can see the KDFs for the delta power spectral density learned from real data. The density shape for the sleep class N3 seems surprising complex (two maxima) but KDFs lead to a significant improvement of accuracy on testing data (i.e. it does not seem to be an overfitting effect).

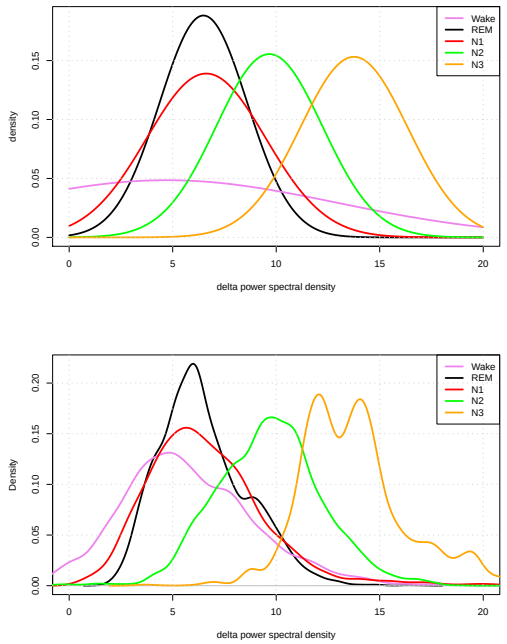


Figure 3: Gaussian density estimates (top) and Kernel Density Functions (bottom) for the delta power spectral density.

3 From Hidden Markov Models to Dynamic Bayesian Networks

At each clock time t, a Hidden Markov Model (HMM) consists of:

- an unobserved state variable Y_t taking a finite number of states. In the sleep analysis the unobserved variable will be the sleep stage and it will take states $\mathcal{Y} = \{Wake, N1, N2, N3, REM\}$ and
- a set of observed variables $X_{i,t}, i = 1, \ldots, k$ (e.g., the spectral density at a given frequency range of the EEG signal).

A new state y_{t+1} is entered based upon a transition probability distribution $P(Y_{t+1} = y'|Y_t = y)$ for $y, y' \in \mathcal{Y}$ which depends on the previous state Y_t (this is called the Markovian property). This allows to exploit the probabilistic dependence of successive sleep stages. Transitions between some of the sleep stages are much more likely than between others. For example, a transition from stage "wake" (W) directly to stage "deep sleep" (N3) is quite unlikely. After each transition is made,

an observation x is produced according to a conditional probability distribution $P(X_{i,t+1} = x|Y_{t+1} = y)$ which depends on the current state y of Y_{t+1} only. In Figure 4 we present and example of a structure of the two consecutive stages of a Hidden Markov Model.

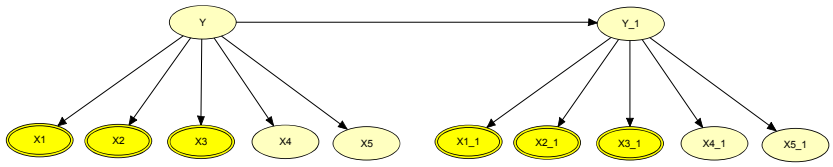


Figure 4: The structure of two consecutive stages of a Hidden Markov Model.

The conditional probability distributions $P(X_{t+1} = x|Y_{t+1} = y)$ are assumed to be stationary (i.e., they do not depend on time t) and can be defined (and learned) in the same way as the conditional probability distributions of the NB model as discussed in Section 2. We used the CDFs due to their better performance in the experiments on real data. The transition probability distribution $P(Y_{t+1} = y'|Y_t = y)$ is discrete and it is also assumed to be stationary. It is easily estimated from training data by normalizing the corresponding contingency table.

As we will see in the Section 4 the HMMs perform better than NB models. However, we conjectured that there is still a room for additional improvement since some of the features should not be treated as dependent on the current state only but also on the previous state. This is because some features are witnesses of state transitions. Therefore we include the model binary variables $Z_{j,t+1}$ that depend on state variables Y_{t+1} and Y_t. These distributions are again assumed to be stationary. They are estimated from training data by normalizing the corresponding contingency table.

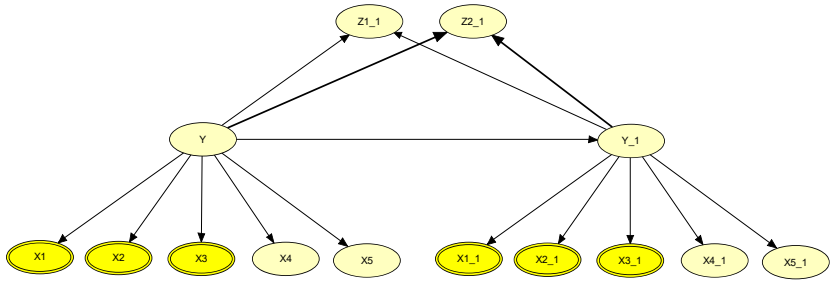


Figure 5: The structure of two consecutive stages of a Dynamic Bayesian Network.

In Figure 5 we present an example of a structure of the two consecutive stages of a Dynamic Bayesian Network. There are two state transition witness features –

denoted Z_1 and Z_2 in the figure. We will refer to them as transition features. Note that the Markovian property does not hold any more since the values of features $X_{i,t+1}$ are not independent of the past given the current state of Y_{t+1}. We can see that if a transition feature $Z_{j,t+1}$ is known a path from the past gets open.

The probabilistic inference in these DBN is very similar to the standard inference in the HMMs, where the Viterbi algorithm [6] is used. The difference is that in each time step we modify the transition matrix representing transition probability distribution $P(Y_{t+1}|Y_t)$ by multiplying it by a probability potential $P(Z_{j,t+1} = z_{j,t+1}|Y_{t+1}, Y_t)$ where $z_{j,t+1}$ is the state of the transition feature observed at time $t + 1$. We do this multiplication for each transition feature. The computational process in time slice t is completed (as it is also done in the standard Viterbi algorithm) by selecting the most probable state y_t for each state y_{t+1} and by normalizing the distribution over the states of Y_{t+1}. Then a backward pass is performed to find a most probable configuration of state variables. See Algorithm 1.

Input: A DBN defined by the conditional probability distributions $P(Y_{t+1}|Y_t)$, $P(Z_{j,t+1}|Y_{t+1}, Y_t)$, $P(X_{i,t+1}|Y_{t+1})$, and $P(Y_1)$, feature evidence:
- $x_{i,t}$ for features $X_{i,t}, i = 1, \ldots, n_X, t = 1, \ldots, N$, and
- $z_{j,t}$ for transition features $Z_{j,t}, j = 1, \ldots, n_Z, t = 2, \ldots, N$.

Output: The most probable state values y_t for $t = 1, \ldots, N$

$S_1 \leftarrow P(Y_1) \cdot \prod_i P(X_{i,1} = x_{i,1}|Y_1)$;
for $t \leftarrow 2$ ***to*** N **do**
  for $y \in \mathcal{Y}$ **do**
    $R \leftarrow S_{t-1} \cdot P(Z_{j,t} = z_{j,t}|Y_t = y, Y_{t-1}) \cdot P(Y_t = y|Y_{t-1})$
     $\cdot P(X_{i,t} = x_{i,t}|Y_t = y)$;
    $T_t(y) \leftarrow \arg\max_{y' \in \mathcal{Y}} R(y')$;
    $S_t(y) \leftarrow \max_{y' \in \mathcal{Y}} R(y')$;
  end
  $S_t \leftarrow \frac{S_t}{\sum_{y'} S_t(y')}$;
end
$y_N \leftarrow \arg\max_{y \in \mathcal{Y}} S_N(y)$;
for $t \leftarrow N$ ***to*** 2 **do**
  $y_{t-1} \leftarrow T_t(y_t)$;
end

Algorithm 1: The inference algorithm for solving the sleep analysis DBN

4 Experiments

We learned our models on a training dataset that consisted of 37 hypnograms. The models were tested on a (testing) dataset that also consisted of 37 (different) hypnograms. Altogether we used 46 features. In the DBN model we used the

Arousal feature as the transition feature while in all other models it was treated as a standard feature related to the current state only. The results of experiments are summarized in Table 1 where the methods are presented in the ascending order given by their accuracy.

Table 1: The average accuracy of the tested methods on testing data.

Method	Accuracy
No Information Rate	41.57%
Naive Bayes with CGDs	49.13%
Philips Respironics	57.43%
Naive Bayes with KDFs	65.06%
Hidden Markov Model	65.71%
Dynamic Bayesian Network	67.02%

In Figure 6 we compare the accuracy of the proposed methods on the testing dataset. Each point in the plot corresponds to one hypnogram from the testing dataset. To see whether the methods differ significantly we performed the Wilcoxon signed rank test for the pairs of methods results presented in Figure 6. We can conclude that:

- The NB model with Kernel Density Functions achieved significantly better accuracy than the NB model with Conditional Gaussian Distributions with the p-value = 5.093e-10
- Hidden Markov Models achieved significantly better accuracy than the NB model with Kernel Density Functions with the p-value = 7.577e-05.
- DBNs achieved significantly better accuracy than HMMs with the p-value = 1.533e-05.
- DBNs achieved significantly better accuracy than Philips Respironics with the p-value = 0.001416

See Figure 7 for the comparisons of the hypnograms predicted by tested methods with the expert for a selected hypnogram of one person. Though, the predicted hypnograms may look very similar, at a closer look, we can see differences that imply different accuracy on the testing dataset. For example, if we have a look at the first two hours of the patient's sleep we can see that the HMM and DBN leads to less oscillations than NB. The DBN further improves the fit by widening the third period of the N2 sleep stage. The DBN also further reduces some oscillations in the latter periods of the sleep. The accuracy for this patient hypnogram was: 77.13% (for NB), 77.53% (for HMM), and 79.98% (for DBN). The Philips Respitronics has the accuracy of 52.50% for this patient.

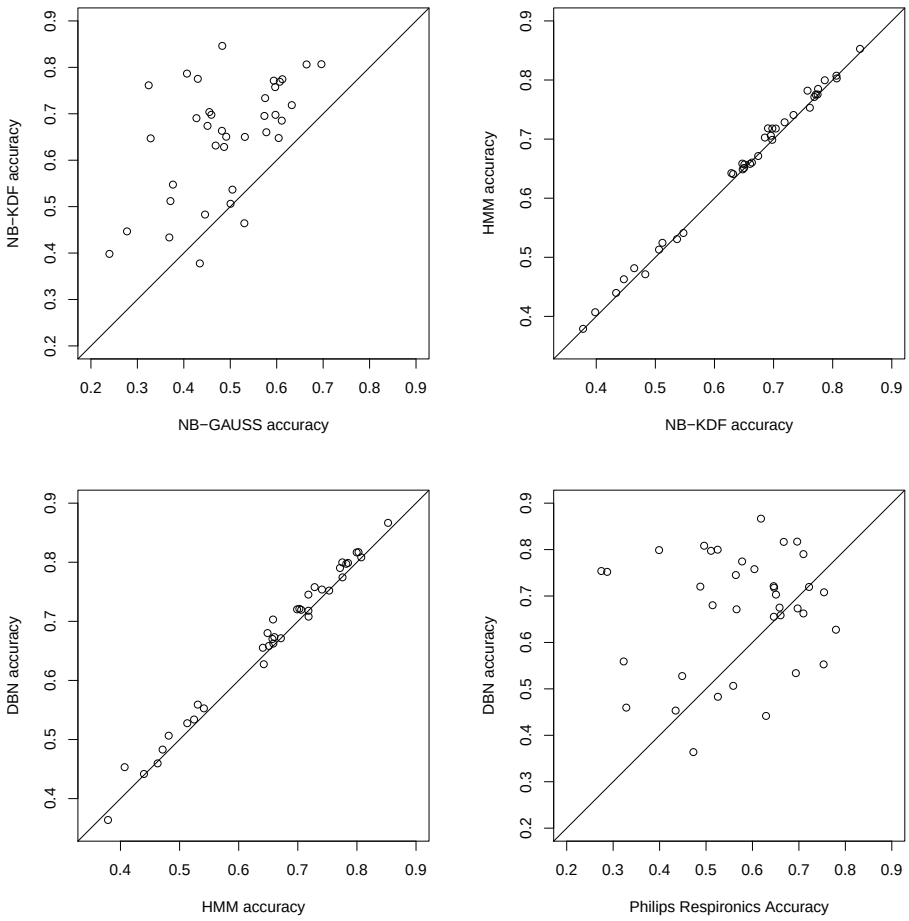


Figure 6: Comparisons of the methods' accuracy on real data.

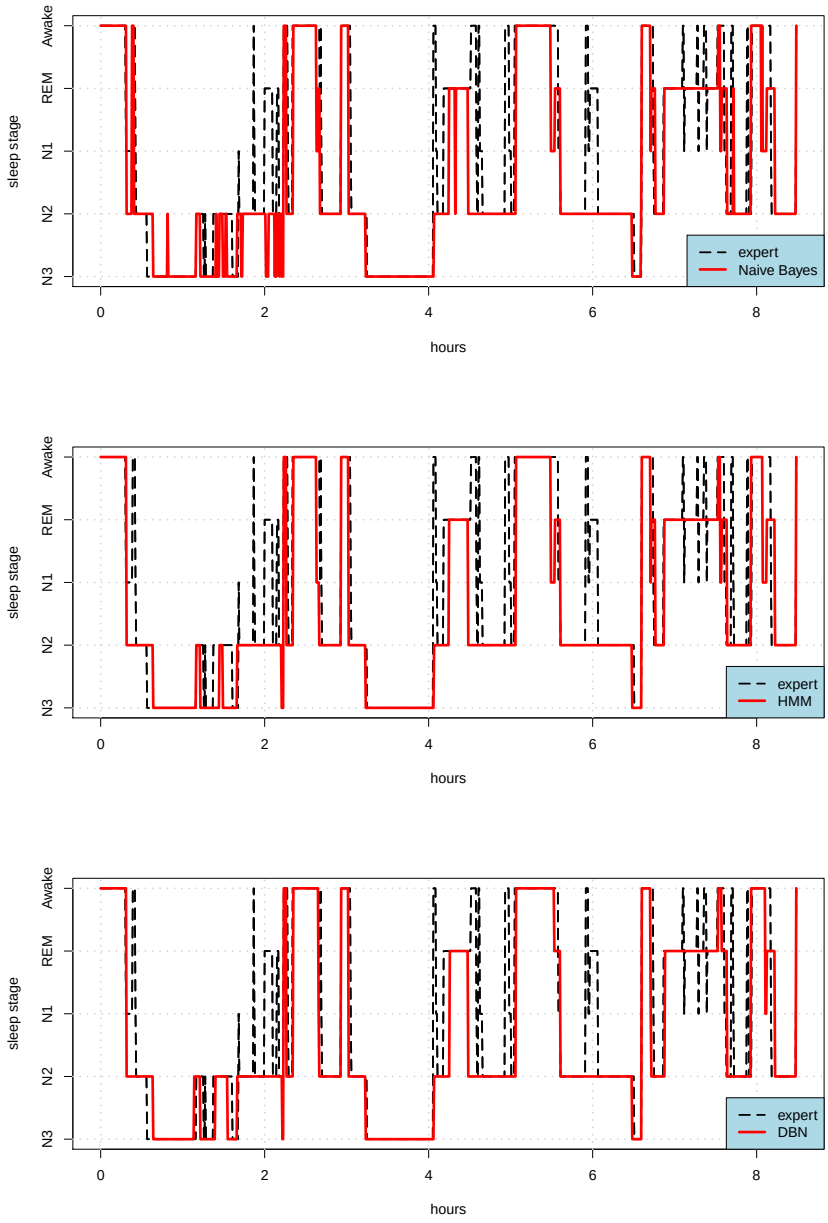


Figure 7: Comparisons of the predicted hypnograms with the expert hypnogram for one selected patient.

5 Conclusions and Future Work

In the problem of the classification of sleep stages Hidden Markov models and Dynamic Bayesian networks have achieved a better accuracy. They also better filter oscillations in hypnograms than Naive Bayes models. It is because they take into their consideration the sleep stage in the previous time step which is closely related to the sleep stage in the current time step. Using the Arousal feature we have demonstrated that the implementation of transition-related features in a Dynamic Bayesian Network helps to further improve the predictions accuracy. Other candidate features of this type are spindles, K-complexes, etc. These were not available in our dataset.

A model we did not use in our experiments but with which we would also like to compare in our future work is a Recurrent Neural Network. We also may consider to include time factors into our models - as are the time spent in a sleep stage and the total time spent in a sleep.

Another future goal is to use the proposed methods for automatic detection of sleep related disorders, which is of great interest of medical doctors. Yet another interesting research direction might be opened by lifting the assumption of five predefined sleep stages and by considering the variable representing the sleep stages as a truly hidden variable with an unknown number of states.

Acknowledgements

This work was supported by the Czech Science Foundation (projects 16-12010S and 17-08182S). This work benefited from a personal communication with Hynek Kružík and Nela Grimová.

References

[1] T. Al-Ani, Y. Hamam, R. Fodil, F. Lofaso, and D. Isabey. Using hidden Markov models for sleep disordered breathing identification. *Simulation Modelling Practice and Theory*, 12:117–128, 2004.

[2] A. Flexer, G. Gruber, and G. Dorffner. A reliable probabilistic sleep stager based on a single EEG signal. *Artificial Intelligence in Medicine*, 33:199–207, 2005.

[3] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan. *The AASM Manual for the Scoring of Sleep and Associated Events.* American Academy of Sleep Medicine, Westchester, Illinois, 2007.

[4] W. D. Penny and S. J. Roberts. Gaussian observation Hidden Markov models for EEG analysis. Technical Report Technical Report, TR-98-12, Imperial College, London, 1998.

[5] A. Rechtschaffen and A. Kales, editors. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Public Health Service, U.S. Government Printing Office, Washington, D.C., 1968.

[6] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

Stochastic Models of Wage Distributions: Empirical Comparison

Michal Vrabec
Dept. of Statistics
University of Economics, Prague
e-mail: vrabec@vse.cz

Petr Berka
Dept. of Information and Knowledge Engineering
University of Economics, Prague
and
Dept. of Informatics and Mathematics
University of Finance and Administration, Prague
e-mail: berka@vse.cz

Abstract

A number of stochastic models for modeling time series data can be found in the literature. Among them models based on Log-normal distribution are more traditional, while models using Johnson SB or Johnson SU distributions were introduced recently. We present basic properties of the above-mentioned distributions and discuss their usability to model economic data. Data concerning the wages of more than two million Czech employees collected for more than twenty years are used for the comparison.

1 Introduction

Statistical analysis of the development of the wage and income distribution is a crucial precondition for economic modeling of the labour market processes. One of the most discussed characteristics of the wage distribution is the average wage. There is an ongoing debate about the suitability of the average as a measure of the wage level. There are proposals to replace the average by median, and/or to consider additional characteristics like variability or percentiles. In our opinion, it is necessary to work with the entire wage distribution.

If the wage distribution is more or less "smooth", it can be adequately modeled with the aid of a suitable theoretic (continuous) distribution, such as a log-normal one ([2]). But as shown e.g. in ([3]), as far as wages are concerned, the log-normal

distribution is not the best-fitting one and this distribution is most often used mainly because of its convenient theoretical qualities. Following this argument, we present an empirical comparison of the log-normal distribution and log-logistic distribution with Johnson SB and Johnson SU distributions.

2 Used distributions

2.1 Log-normal Distribution

Log-normal distribution (sometimes also called Galton distribution) is a continuous probability distribution of a random variable whose logarithm is normally distributed. The formula (1) represents the density of a three-parameter log-normal distribution: here μ is the location parameter and σ is the scale parameter ($\sigma > 0$) for the normally distributed logarithm $\ln(X)$.

$$p(x) = \frac{1}{(x-\gamma)\sigma\sqrt{2\pi}} \times \exp\left[-\frac{(\ln(x-\gamma)-\mu)^2}{2\sigma^2}\right], \ x \geq 0 \tag{1}$$

If $\gamma = 0$ then the three-parameter log-normal distribution changes into two-parameter one, as shown in formula (2).

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \times \exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right], \ x \geq 0 \tag{2}$$

2.2 Log-logistic Distribution

Log-logistic distribution is the probability distribution of a random variable whose logarithm has a logistic distribution. The formula (3) represents the density of this distribution: here γ is the location parameter ($\gamma = 0$ yields a two-parameter distribution) and α is shape parameter ($\alpha > 0$) and β is scale parameter ($\beta > 0$).

$$p(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2} \tag{3}$$

2.3 Johnson Distribution

Johnson distributions [1] are based on a transformation of the standard normal variable. Given a continuous random variable X whose distribution is unknown and is to be approximated, Johnson proposed three normalizing transformations having the general form:

$$Z = \gamma + \delta f\left(\frac{X-\mu}{\sigma}\right) \tag{4}$$

where $f(.)$ denotes the transformation function, Z is a standard normal random variable, γ and δ are shape parameters, σ is a scale parameter and μ is a location parameter. Without loss of generality, it is assumed that $\delta > 0$ and $\sigma > 0$. Johnson distributions include three forms: log-normal, bounded and unbounded.

The most simple transformation defines the log-normal system of distributions denoted by S_L

$$Z = \gamma + \delta \ln\left(\frac{X-\mu}{\sigma}\right), \ X > \Theta \tag{5}$$

The bounded system of distributions S_B is defined by

$$Z = \gamma + \delta \ln\left(\frac{X-\mu}{\mu+\sigma-X}\right), \ \mu < X < \mu + \sigma \tag{6}$$

S_B curves cover bounded distributions. The distributions can be bounded on the lower end, the upper end or both ends. This family covers Gamma distributions, Beta distributions and many others.

The unbounded system of distributions S_U is defined by

$$Z = \gamma + \delta \ln\left\{\left(\frac{X-\mu}{\sigma}\right) + \left[\left(\frac{X-\mu}{\sigma}\right)^2 + 1\right]^{1/2}\right\}, \ -\infty < X < \infty \tag{7}$$

The S_U curves are unbounded and cover the t and normal distributions, among others.

Using the fact that, after the transformation in (4), Z follows standard normal distribution, the probability density function $p(y)$ of each of the family in the Johnson system can be derived. If X follows the Johnson distribution and $Y = \frac{X-\mu}{\sigma}$ then, for S_L family,

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1}{y} \times \exp\left[-\frac{1}{2}\left(\gamma + \delta \ln(y)\right)^2\right] \ \mu < x < \infty \tag{8}$$

similarly, for the S_B family,

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1-y}{y} \times \exp\left[-\frac{1}{2}\left(\gamma + \delta \ln\left(\frac{y}{1-y}\right)\right)^2\right] \ \mu < x < \mu + \sigma \tag{9}$$

and for the S_U family,

$$p(y) = \frac{\delta}{\sqrt{2\pi}} \times \frac{1}{\sqrt{y^2+1}} \times \exp\left[-\frac{1}{2}\left(\gamma + \delta \ln\left(y + \sqrt{y^2+1}\right)\right)\right] -\infty < x < \infty \tag{10}$$

In general, the the probability density function of X is given by

$$p(x) = \frac{\delta}{\sqrt{2\pi}} \times f'\left(\frac{x-\mu}{\sigma}\right) \times \exp\left[-\frac{1}{2}\left(\gamma + \delta f\left(\frac{x-\mu}{\sigma}\right)\right)^2\right] \quad x \in G \tag{11}$$

where

$$f'(y) = \begin{cases} \frac{1}{y}, & \text{for the } S_L \text{ family} \\ \frac{1}{y(1-y)}, & \text{for the } S_B \text{ family} \\ \frac{1}{\sqrt{y^2+1}}, & \text{for the } S_U \text{ family} \end{cases} \tag{12}$$

and

$$f(y) = \begin{cases} \ln(y), & \text{for the } S_L \text{ family} \\ \ln\left(\frac{y}{1-y}\right), & \text{for the } S_B \text{ family} \\ \ln\left(y + \sqrt{y^2+1}\right), & \text{for the } S_U \text{ family} \end{cases} \tag{13}$$

The support G of the distribution is:

$$G = \begin{cases} \langle \mu; \infty) , & \text{for the } S_L \text{ family} \\ \langle \mu; \mu + \sigma \rangle , & \text{for the } S_B \text{ family} \\ (-\infty; \infty), & \text{for the } S_U \text{ family} \end{cases} \tag{14}$$

3 Modeling wage distributions

3.1 Used data

We work with time series of wages in Czech Republic over the years 1995 - 2017. The annual data are reported in quarterly units; our study observes the average wages in the second quarter of each year. The scope of the data set on which the analyses were carried out was gradually increased from more than 300,000 observations in 1995 to more than two million in 2017. This data is structured in a very detailed way. The wage values are divided into intervals with widths of 500 CZK. Such a detailed structure enables us to achieve quite accurate results. We have basic characteristics of wages in the entire period at our disposal. The analysis was aimed at creating a model for probability distribution of wages (estimating the parameters of the probability density). We used only the data for the years 2015-2017 in the experiments. Table 1 shows basic characteristics of the data, all numbers except the sample size are in Czech crowns (CZK).

3.2 Parameter estimation

We used the SAS system and EasyFit program for computations. Figures (Fig. 1), (Fig. 2), (Fig. 3) and (Fig. 4) show the fit of the distributions on data

characteristics	year 2015	year 2016	year 2017
sample size	2 098 854	2 119 396	2 185 573
average wage	26 369	27 668	29 166
standard deviation	19 903	20 478	20 749
10th percentile	12 978	13 944	14 982
lower quartile	17 290	18 391	19 547
median	22 658	23 757	25 135
upper quartile	29 566	30 963	32 610
90th percentile	40 162	42 026	44 334
modus	8 635	9 275	10 296

Table 1: Basic characteristics of the used data

from the year 2015. We also performed the Kolmogorov-Smirnov test to assess the quality of the model. We tested the null hypothesis "H0: the data follow the specified distribution" against the alternative hypothesis "H1: the data do not follow the specified distribution" Table 2 gives the results of this test in terms of the Kolmogorov-Smirnov statistics and the rank of the model (in both cases, the lower is the value the better is the model).

distribution	year 2015		year 2016		year 2017	
	statistics	rank	statistics	rank	statistics	rank
2 par. log-normal	0,03808	3	0,03877	3	0,03949	3
3 par. log-normal	0,03739	2	0,03809	2	0,03886	2
log-logistic	0,01839	1	0,01667	1	0,01732	1
Johnson SB	0,06982	4	0,06605	4	0,0621	4

Table 2: Results of the Kolmogorov-Smirnov test

4 Conclusions

The aim of the analysis was to compare several models of probability distribution of wags in Czech Republic. The experiments show that the best model is the log-logistic distribution with three parameters. This confirms the previously achieved results ([4]). Anyway, as the wage variability grows over the years and empirical density's curves became less smooth, mixture models have potential to provide better models of wage distributions in the future. And good models that are able to make good predictions of the future wage distributions are necessary for various socio-economic considerations.

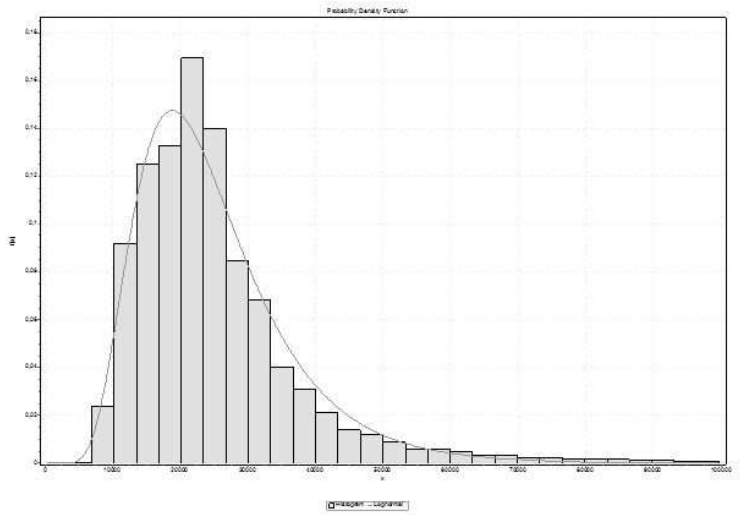


Figure 1: Two parameters log-normal distribution for wages from the year 2015. Here $\mu = 10,032$ and $\sigma = 0,43343$.

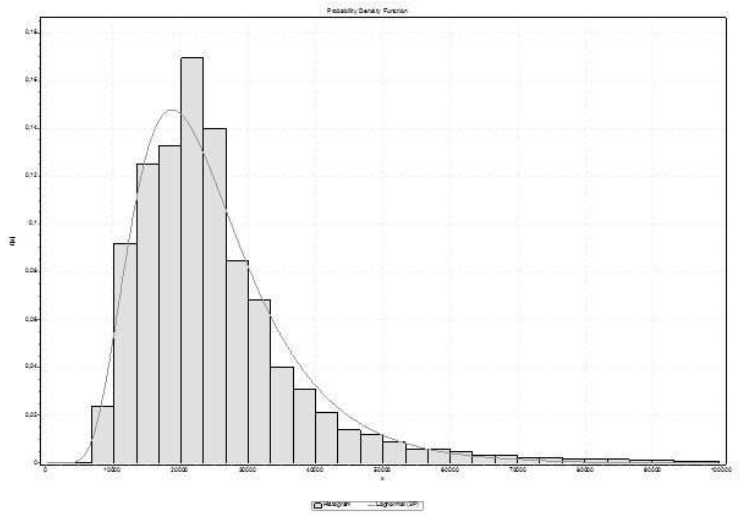


Figure 2: Three parameters log-normal distribution for wages from the year 2015. Here $\mu = 10,02$, $\sigma = 0,43841$ and $\gamma = 250$.

References

[1] N. L. Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, 36(3/4):297–304, 1949.

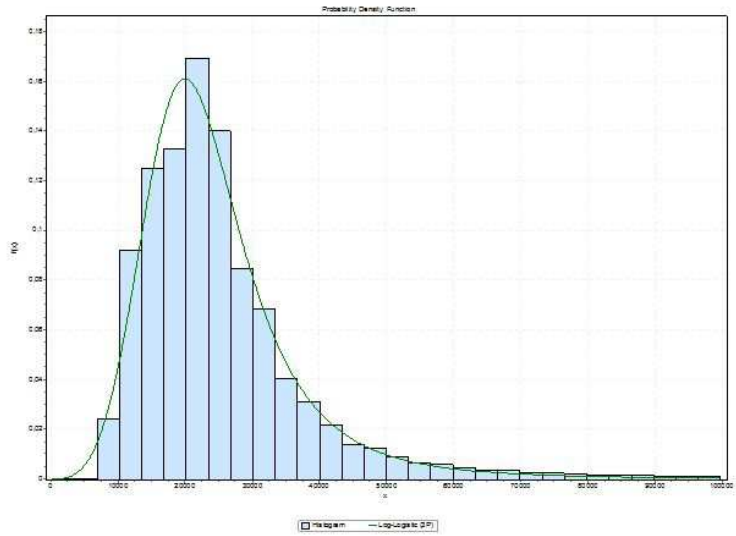


Figure 3: Log-logisic distribution for wages from the year 2015.

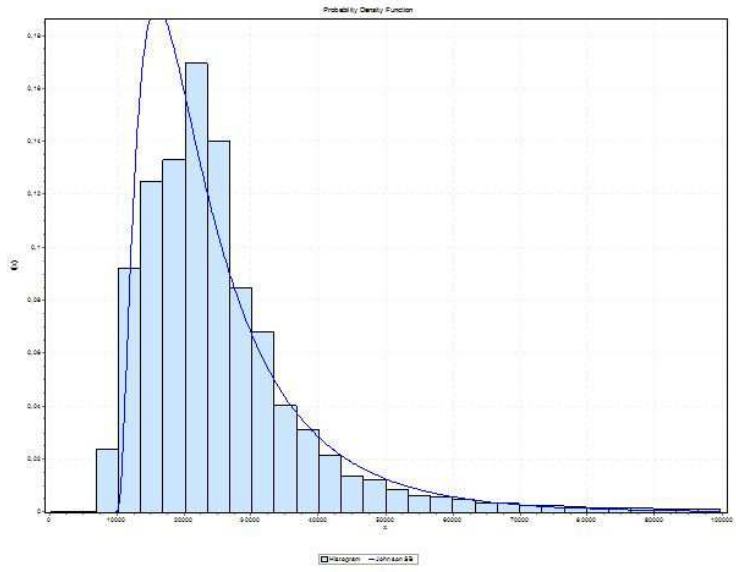


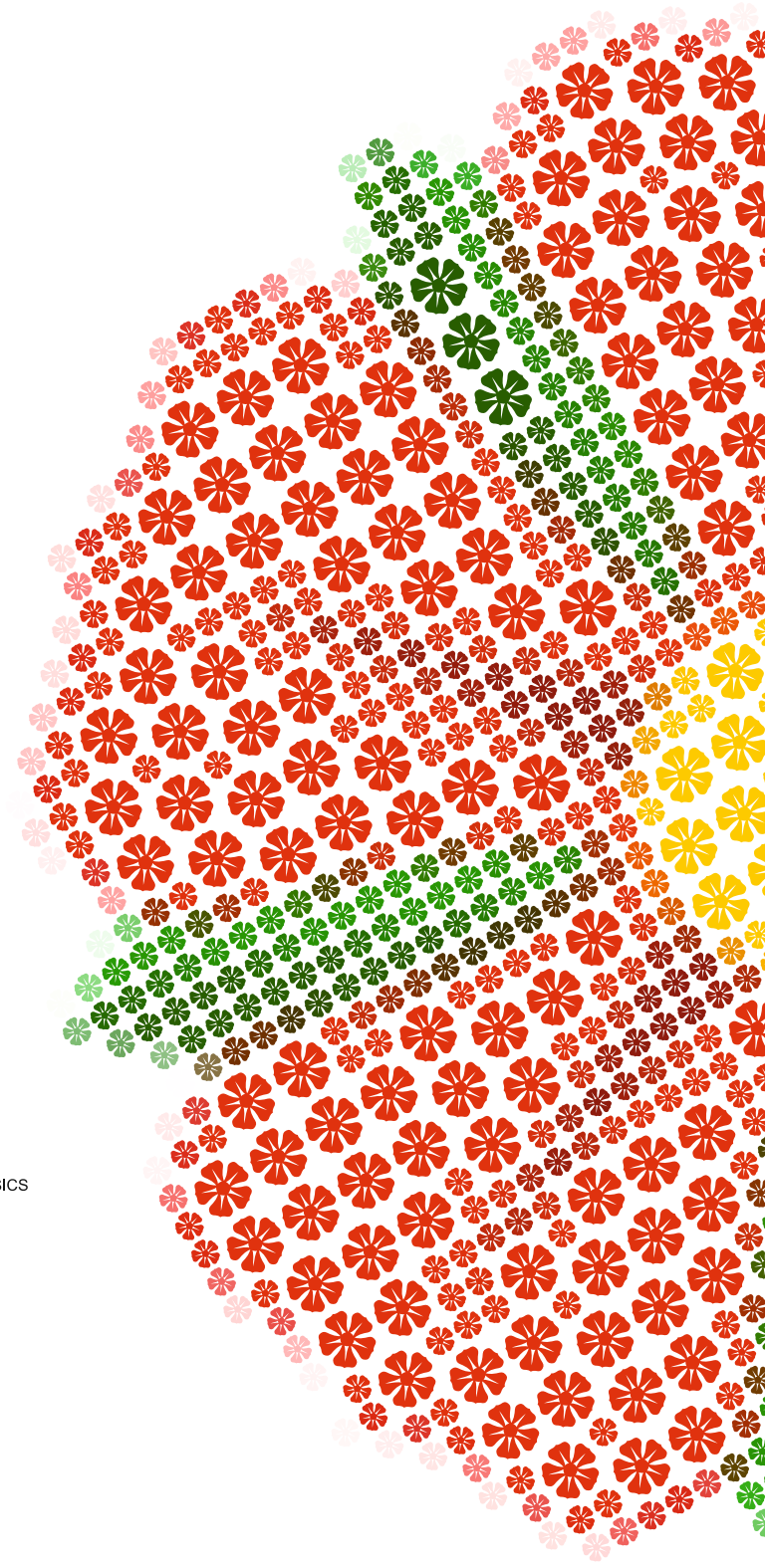
Figure 4: Johnson SB distribution for wages from the year 2015. Here $\mu = 9818,9$, $\sigma = 1,8883E+5$, $\gamma = 3,1927$ and $\delta = 1,1751$.

[2] L. Marek and M. Vrabec. Forecast of the income distribution in the czech republic in 2011. In *International Conference on Applied Business Research ICABR 2010*, page 142, 2010.

[3] M. Matějka and K. Duspivová. The czech wage distribution and the minimum wage impacts: an empirical analysis. *Statistika*, 93(2):61–75, 2013.

[4] M. Vrabec and L. Marek. Model for distribution of wages. In *Proceedings of the Applications of Mathematics and Statistics in Economics AMSE 2016*, pages 378–386, Banska Bystrica, 2016.

List of Authors



matfyzpress

PUBLISHING HOUSE
OF THE FACULTY OF MATHEMATICS AND PHYSICS
CHARLES UNIVERSITY IN PRAGUE

ISBN 978-80-7378-361-7