Advances in Intelligent Systems and Computing 456

Maria Brigida Ferraro
Paolo Giordani
Barbara Vantaggi
Marek Gagolewski
María Ángeles Gil
Przemysław Grzegorzewski
Olgierd Hryniewicz *Editors*

# Soft Methods for Data Science

Springer

# Advances in Intelligent Systems and Computing

Volume 456

**Series editor**

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

## *About this Series*

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within “Advances in Intelligent Systems and Computing” are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

## *Advisory Board*

More information about this series at http://www.springer.com/series/11156

Maria Brigida Ferraro · Paolo Giordani
Barbara Vantaggi · Marek Gagolewski
María Ángeles Gil · Przemysław Grzegorzewski
Olgierd Hryniewicz
Editors

# Soft Methods for Data Science

Springer

*Editors*
Maria Brigida Ferraro
Department of Statistical Sciences
Sapienza University of Rome
Rome
Italy

Paolo Giordani
Department of Statistical Sciences
Sapienza University of Rome
Rome
Italy

Barbara Vantaggi
Department of Basic and Applied Sciences for Engineering
Sapienza University of Rome
Rome
Italy

Marek Gagolewski
Department of Stochastic Methods, Systems Research Institute
Polish Academy of Sciences
Warsaw
Poland

María Ángeles Gil
Department of Statistics and Operational Research and Mathematics Didactics
University of Oviedo
Oviedo
Spain

Przemysław Grzegorzewski
Department of Stochastic Methods, Systems Research Institute
Polish Academy of Sciences
Warsaw
Poland

Olgierd Hryniewicz
Department of Stochastic Methods, Systems Research Institute
Polish Academy of Sciences
Warsaw
Poland

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-42971-7 ISBN 978-3-319-42972-4 (eBook)
DOI 10.1007/978-3-319-42972-4

Library of Congress Control Number: 2016945947

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

# Preface

This volume is a collection of peer-reviewed papers presented at the 8th International Conference on Soft Methods in Probability and Statistics—SMPS 2016, held in Rome (Italy) during September 12–14, 2016. The series of biannual international conferences on Soft Methods in Probability and Statistics (SMPS) started in Warsaw in 2002. Subsequent events in this series took place in Oviedo (2004), Bristol (2006), Toulouse (2008), Oviedo/Mieres (2010), Konstanz (2012), and Warsaw (2014). SMPS 2016 was organized by the Department of Basic and Applied Sciences for Engineering and the Department of Statistical Sciences, Sapienza University of Rome, Italy.

Over the last 50 years in different areas such as decision theory, information processing, and data mining, the interest to extend probability theory and statistics has grown. The common feature of those attempts is to widen frameworks for representing different kinds of uncertainty: randomness, imprecision, vagueness, and ignorance. The scope is to develop more flexible methods to analyze data and extract knowledge from them. The extension of classical methods consists in "softening" them by means of new approaches involving fuzzy set theory, possibility theory, rough sets, or having their origin in probability theory itself, like imprecise probabilities, belief functions, and fuzzy random variables.

Data science aims at developing automated methods to analyze massive amounts of data and extract knowledge from them. In the recent years the production of data is dramatically increasing. Every day a huge amount of data coming from everywhere is collected: mobile sensors, sophisticated instruments, transactions, Web logs, and so forth. This trend is expected to accelerate in the near future. Data science employs various programming techniques and methods of data wrangling, data visualization, machine learning, and probability and statistics. The soft methods proposed in this volume represent a suit of tools in these fields that can also be useful for data science.

The volume contains 65 selected contributions devoted to the foundation of uncertainty theories such as probability, imprecise probability, possibility theory, soft methods for probability and statistics. Some of them are focused on robustness, non-precise data, dependence models with fuzzy sets, clustering, mathematical models for decision theory and finance.

We would like to thank all contributing authors, organizers of special sessions, program committee members, and additional referees who made it possible to put together the attractive program of the conference. We are very grateful to the plenary speakers, Ana Colubi (University of Oviedo, Spain), Thierry Denoeux (University of Technology of Compiégne, France), Massimo Marinacci (Bocconi University, Italy) for their very interesting talks: “On some functional characterizations of (fuzzy) set-valued random elements”, “Beyond Fuzzy, Possibilistic and Rough: An Investigation of Belief Functions in Clustering” and “A non Bayesian Approach to Measurement Problems”, respectively. We would like to express our gratitude also to INDAM-GNAMPA for the financial support. Furthermore, we would like to thank the editor of the Springer series Advances in Soft Computing, Prof. Janusz Kacprzyk, and Springer-Verlag for the dedication to the production of this volume.

Rome
June 2016

Maria Brigida Ferraro
Paolo Giordani
Barbara Vantaggi
Marek Gagolewski
María Ángeles Gil
Przemysław Grzegorzewski
Olgierd Hryniewicz

# Organization

## General Chairs

Maria Brigida Ferraro, Rome, Italy
Paolo Giordani, Rome, Italy
Barbara Vantaggi, Rome, Italy

## Executive Board

María Ángeles Gil, Oviedo, Spain
Przemysław Grzegorzewski, Warsaw, Poland
Olgierd Hryniewicz, Warsaw, Poland

## Program Committee

Bernard de Baets, Ghent, Belgium
Christian Borgelt, Mieres, Spain
Giulianella Coletti, Perugia, Italy
Ana Colubi, Oviedo, Spain
Ines Couso, Oviedo, Spain
Gert De Cooman, Ghent, Belgium
Thierry Denoeux, Compiégne, France
Didier Dubois, Toulouse, France
Fabrizio Durante, Bolzano, Italy
Pierpaolo D'Urso, Rome, Italy
Gisella Facchinetti, Lecce, Italy
Peter Filzmoser, Vienna, Austria
Lluís Godo, Barcelona, Spain
Gil González Rodríguez, Oviedo, Spain
Mario Guarracino, Naples, Italy

Janusz Kacprzyk, Warsaw, Poland
Frank Klawonn, Braunschweig/Wolfenbüttel, Germany
Rudolf Kruse, Magdeburg, Germany
Massimo Marinacci, Milan, Italy
Radko Mesiar, Bratislava, Slovakia
Enrique Miranda, Oviedo, Spain
Martin Stepnicka, Ostrava, Czech Republic
Wolfgang Trutschnig, Salzburg, Austria
Stefan Van Aelst, Ghent, Belgium

## Publication Chair

Marek Gagolewski, Warsaw, Poland

## Special Session Organizers

Patrizia Berti, Modena, Italy
Angela Blanco-Fernández, Oviedo, Italy
Andrea Capotorti, Perugia, Italy
Fabrizio Durante, Bolzano, Italy
Luis Angel García-Escudero, Valladolid, Spain
Brunero Liseo, Rome, Italy
Agustin Mayo-Iscar, Valladolid, Spain
Enrique Miranda, Oviedo, Spain
Niki Pfeifer, Munich, Germany
Ana B. Ramos-Guajardo, Oviedo, Spain
Giuseppe Sanfilippo, Palermo, Italy
Beatriz Sinova, Oviedo, Spain
Pedro Terán, Oviedo, Spain
Wolfgang Trutschnig, Salzburg, Austria

## Additional Reviewers

Serena Arima, Salem Benferhat, Patrizia Berti, Angela Blanco-Fernández, Christian Borgelt, Andrea Capotorti, Marco Cattaneo, Jasper de Bock, Alessandro De Gregorio, Sara De la Rosa de Sáa, Marco Di Zio, Marta Disegna, Scott Ferson, Gianna Figá-Talamanca, Enrico Foscolo, Marek Gagolewski, Luis Angel García-Escudero, José Luis García-Lapresta, Angelo Gilio, Rita Giuliano, Stefania Gubbiotti, Maria Letizia Guerra, Radim Jiroušek, Tomas Kroupa, Brunero Liseo, Massimo Marinacci, Agustin Mayo-Iscar, Enrique Miranda, Piotr Nowak, David Over, Sonia Pérez-Fernández, Davide Petturiti, Niki Pfeifer, Ana B. Ramos-Guajardo, Pietro Rigo, Giuseppe Sanfilippo, Fabian Schmidt, Rudolf Seising, Beatriz Sinova, Damjan Skulj, Yann Soullard, Andrea Tancredi, Luca Tardella, Pedro Terán, Matthias Troffaes, Tiziana Tuoto, Arthur Van Camp, Paolo Vicig.

## Local Organization

Marco Brandi
Massimiliano Coppola—conference website chair
Francesco Dotto
Liana Francescangeli—administrative staff
Tullia Padellini
Alessandra Pelorosso—administrative chair
Marco Stefanucci

# Contents

# Mean Value and Variance of Fuzzy Numbers with Non-continuous Membership Functions

**Luca Anzilli and Gisella Facchinetti**

**Abstract** We propose a definition of mean value and variance for fuzzy numbers whose membership functions are upper-semicontinuous but are not necessarily continuous. Our proposal uses the total variation of bounded variation functions.

## 1 Introduction

In this work we face the problem of defining the mean value and variance of fuzzy numbers whose membership functions are upper-semicontinuous but are not necessarily continuous. The literature presents a high number of definitions in the continuous case [3–6, 9] also because the average has often been counted among the ranking modes of fuzzy numbers. The starting point is that a fuzzy number is defined by a membership function that is of bounded variation. Even in this more general context, it is possible to introduce a weighted average by means of a classical variational formulation and by $\alpha$-cuts, as well as many authors do in the continuous case. In the non-continuous case, we introduce the lower and upper weighted mean values, but the generality of the weights doesn't let to obtain the weighted mean value as the middle point of the previous ones. This property will be obtained either in the continuous case or for particular weight functions. An interesting property of this new version is connected to the view of the weighted mean value in a possibilistic framework, as in the continuous case Carlsson and Fullér [3] and Fullér and Majlender [5] do. This property is interesting and harbinger of future developments also in the non-continuous case. Following the same line, we pass to introduce the concept of variance and we suggest two different definitions as happens in the case of continuous membership functions.

L. Anzilli (✉) · G. Facchinetti
Department of Management, Economics, Mathematics and Statistics,
University of Salento, Lecce, Italy
e-mail: luca.anzilli@unisalento.it

G. Facchinetti
e-mail: gisella.facchinetti@unisalento.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_1

## 2 Bounded Variation Functions

In this section we recall some basic properties of functions of bounded variation. For more details see, e.g., [7].

Let $I \subset \mathbb{R}$ be an interval and $u : I \to \mathbb{R}$ be a function. The *total variation* of $u$ on $I$ is defined as $\mathcal{V}_I[u] = \sup \sum_{i=1}^{n} |u(x_i) - u(x_{i-1})|$, where the supremum is taken over all finite partitions $\{x_0, x_1, \ldots, x_n\} \subset I$ with $x_0 < x_1 < \cdots < x_n$. We say that $u$ is a *bounded variation function* on $I$ if $\mathcal{V}_I[u] < +\infty$. We denote $BV(I)$ the space of all bounded variation functions on $I$. The following properties hold: if $u$ is monotone and bounded on $I$ then $u \in BV(I)$ and $\mathcal{V}_I[u] = \sup_I u - \inf_I u$; if $u_1, u_2 \in BV(I)$ and $k$ is a constant then $ku_1, u_1 + u_2, u_1 - u_2, u_1u_2$ belong to $BV(I)$.
Let $u : \mathbb{R} \to \mathbb{R}$ be a bounded variation function. The *total variation function* of $u$ is the increasing function $v_u$ defined by $v_u(x) = \mathcal{V}_{-\infty}^{x}[u]$, where $\mathcal{V}_{-\infty}^{x}[u]$ denotes the total variation of $u$ on $]-\infty, x]$.
The *total variation measure* $|Du|$ of $u$ is defined as the Lebesgue-Stieltjes measure $dv_u$ associated to $v_u$. The positive and negative variations of $u$ are defined, respectively, by the increasing functions

$$u^+(x) = (v_u(x) + u(x))/2, \qquad u^-(x) = (v_u(x) - u(x))/2. \tag{1}$$

We have $u = u^+ - u^-$, the so-called Jordan decomposition of $u$. The integral of a measurable function $g$ with respect to $|Du|$ is given by

$$\int g(x)\,|Du| = \int g(x)\,du^+(x) + \int g(x)\,du^-(x) \tag{2}$$

where $du^+$ and $du^-$ are the Lebesgue-Stieltjes measures associated to $u^+$ and $u^-$, respectively. In general, the total variation measure $|Du|$ is not absolutely continuous with respect to Lebesgue measure. However, if $u$ is an absolutely continuous function then $\int g(x)\,|Du| = \int g(x)\,|u'(x)|dx$.

## 3 Weighted Mean Value of Fuzzy Numbers

In this section we propose a definition of $f$-weighted mean value for a fuzzy number whose membership function is upper-semicontinuous but not necessarily continuous. These properties produce that the fuzzy number has membership function of bounded variation and so we realize to introduce its weighted mean value by its total variation measure.

A fuzzy number $A$ is a fuzzy set of $\mathbb{R}$ with a normal, (fuzzy) convex and upper-semicontinuous membership function $\mu : \mathbb{R} \to [0, 1]$ of bounded support [2]. From the previous definition there exist two functions $\mu_L, \mu_R : \mathbb{R} \to [0, 1]$, where $\mu_L$ is nondecreasing and right-continuous and $\mu_R$ is nonincreasing and left-continuous, such that

$$\mu(x) = \begin{cases} 0 & x < a_1 \text{ or } x > a_4 \\ \mu_L(x) & a_1 \leq x < a_2 \\ 1 & a_2 \leq x \leq a_3 \\ \mu_R(x) & a_3 < x \leq a_4 \end{cases}$$

with $a_1 \leq a_2 \leq a_3 \leq a_4$. The $\alpha$-cut of a fuzzy set $A$, $0 \leq \alpha \leq 1$, is the crisp set $A_\alpha = \{x \in X; \mu(x) \geq \alpha\}$ if $0 < \alpha \leq 1$ and $A_0 = [a_1, a_4]$ if $\alpha = 0$. Every $\alpha$-cut of a fuzzy number is a closed interval $A_\alpha = [a_L(\alpha), a_R(\alpha)]$.

We observe that the membership function $\mu$ of a fuzzy number $A$ is a bounded variation function. We propose a definition of mean value of $A$ using the total variation measure $|D\mu|$ of $\mu$. Let $f$ be a weighting function such that $f > 0$ and $\int_0^1 f(\alpha)\, d\alpha = 1$.

**Definition 1** We define the $f$-weighted mean value of a fuzzy number $A$ as

$$E(A; f) = \int_{-\infty}^{+\infty} x\, f(\mu(x))\, |D\mu| \Big/ \int_{-\infty}^{+\infty} f(\mu(x))\, |D\mu| \,. \tag{3}$$

In order to define the lower and upper mean values of $A$, we consider the positive and negative variations of $\mu$, as defined in (1), given by, respectively,

$$\mu^+(x) = \begin{cases} 0 & x < a_1 \\ \mu_L(x) & a_1 \leq x < a_2 \\ 1 & x \geq a_2 \,, \end{cases} \qquad \mu^-(x) = \begin{cases} 0 & x \leq a_3 \\ 1 - \mu_R(x) & a_3 < x \leq a_4 \\ 1 & x > a_4 \,. \end{cases} \tag{4}$$

We observe that $\mu^+$ is an increasing and right continuous function and that $\mu^-$ is an increasing and left continuous function.

**Definition 2** We define the lower and upper $f$-weighted mean values of $A$, respectively, as

$$E_*(A; f) = \frac{\int_{-\infty}^{+\infty} x\, f(\mu(x))\, d\mu^+(x)}{\int_{-\infty}^{+\infty} f(\mu(x))\, d\mu^+(x)}\,, \quad E^*(A; f) = \frac{\int_{-\infty}^{+\infty} x\, f(\mu(x))\, d\mu^-(x)}{\int_{-\infty}^{+\infty} f(\mu(x))\, d\mu^-(x)}\,. \tag{5}$$

From previous definition and (2) we deduce the following result

**Proposition 1** *We have* $E(A; f) = (1 - w)E_*(A; f) + wE^*(A; f)$, *where* $w \in (0, 1)$ *is defined as* $w = \int_{-\infty}^{+\infty} f(\mu(x))\, d\mu^-(x) \Big/ \int_{-\infty}^{+\infty} f(\mu(x))\, |D\mu|$.

### 3.1 Weighted Mean Value Using $\alpha$-cuts

Following the line present in the classical case of continuous fuzzy numbers, we face the problem to give an expression of lower and upper $f$-weighted mean values using $\alpha$-cuts. This different version lets the possibility to write the previous definitions by the left and right extreme points of fuzzy number $\alpha$-cuts, exactly as happens in the continuous case.

**Proposition 2**

$$E_*(A; f) = \frac{\int_0^1 a_L(\alpha)\, f(\mu(a_L(\alpha)))\, d\alpha}{\int_0^1 f(\mu(a_L(\alpha)))\, d\alpha}, \quad E^*(A; f) = \frac{\int_0^1 a_R(\alpha)\, f(\mu(a_R(\alpha)))\, d\alpha}{\int_0^1 f(\mu(a_R(\alpha)))\, d\alpha}. \tag{6}$$

*Proof* First, we prove the following equalities:

$$\int_{-\infty}^{+\infty} g(x)\, f(\mu(x))\, d\mu^+(x) = \int_0^1 g(a_L(\alpha))\, f(\mu(a_L(\alpha)))\, d\alpha\,, \tag{7}$$

$$\int_{-\infty}^{+\infty} g(x)\, f(\mu(x))\, d\mu^-(x) = \int_0^1 g(a_R(\alpha))\, f(\mu(a_R(\alpha)))\, d\alpha\,. \tag{8}$$

Since $a_L(\alpha) = \min\{x \geq a_1;\ \mu_L(x) \geq \alpha\}$, $\alpha \in [0, 1]$, applying the change of variable formula [7, Theorem 5.42] we obtain $\int_{-\infty}^{+\infty} g(x)\, f(\mu(x))\, d\mu^+(x) = \int_{a_1}^{a_2} g(x)$ $f(\mu(x))\, d\mu_L(x) = \int_0^1 g(a_L(\alpha))\, f(\mu(a_L(\alpha)))\, d\alpha$ and thus (7). We now prove (8). We have $\int_{-\infty}^{+\infty} g(x)\, f(\mu(x))\, d\mu^-(x) = \int_{a_4}^{a_3} g(x)\, f(\mu(x))\, d\mu_R(x) = \int_{-a_4}^{-a_3} g(-y)$ $f(\mu(-y))\, d\mu_R(-y) = \int_{-a_4}^{-a_3} g(-y)\, f(\mu(-y))\, dk(y)$
where we have used the change of variable $y = -x$ and let $k(x) = \mu_R(-x)$. Since $a_R(\alpha) = \max\{x;\ x \leq a_4\,, \mu_R(x) \geq \alpha\}$, $\alpha \in [0, 1]$, letting $h(\alpha) = -a_R(\alpha) = \min\{y;\ y \geq -a_4\,,\ k(y) \geq \alpha\}$ and using the change of variable formula [7, Theorem 5.42] (observing that $k$ is increasing), we have, continuing the above chain of equalities,
$= \int_{k(-a_4)}^{k(-a_3)} g(-h(\alpha))\, f(\mu(-h(\alpha)))\, d\alpha = \int_{\mu_R(a_4)}^{\mu_R(a_3)} g(a_R(\alpha))\, f(\mu(a_R(\alpha)))\, d\alpha =$
$\int_0^1 g(a_R(\alpha))\, f(\mu(a_R(\alpha)))\, d\alpha$. Then (8) is proved. Substituting (7) and (8) in (5) with $g(x) = x$ and $g(x) = 1$ we obtain for $E_*(A; f)$ and $E^*(A; f)$ the expressions given in (6). □

We observe that $w$ of Proposition 1 can be expressed in terms of $\alpha$-cuts as $w = \int_0^1 f(\mu(a_R(\alpha)))\, d\alpha \Big/ \Big(\int_0^1 f(\mu(a_L(\alpha)))\, d\alpha + \int_0^1 f(\mu(a_R(\alpha)))\, d\alpha\Big)$.

### 3.2 The Case $f(\alpha) = 1$

We now consider the special case $f(\alpha) = 1$. We denote by $E(A)$, $E_*(A)$ and $E^*(A)$, respectively, the mean value and lower and upper mean values of a fuzzy number $A$ computed using $f(\alpha) = 1$. This particular case is interesting as we may show that a sufficient condition so that the weight $w$ is equal to $1/2$ is that $f(\alpha) = 1$. This condition is not necessary as we will show later in an example.

**Proposition 3** *We have* $E(A) = \frac{1}{2}\int_{-\infty}^{+\infty} x\,|D\mu| = \frac{1}{2}\int_0^1 (a_L(\alpha) + a_R(\alpha))\,d\alpha$, $E_*(A) = \int_{-\infty}^{+\infty} x\,d\mu^+(x) = \int_0^1 a_L(\alpha)\,d\alpha$ *and* $E^*(A) = \int_{-\infty}^{+\infty} x\,d\mu^-(x) = \int_0^1 a_R(\alpha)\,d\alpha$. *In particular* $E(A) = (E_*(A) + E^*(A))/2$.

*Proof* Since $\mu^+(x)$ is increasing, we have (see Sect. 2) $\int_{-\infty}^{+\infty} d\mu^+(x) = \mathcal{V}_{-\infty}^{+\infty}[\mu^+] = \sup\mu^+ - \inf\mu^+ = 1$, where last equality follows from (4). Similarly we get $\int_{-\infty}^{+\infty} d\mu^-(x) = 1$. Then, from (2), $\int_{-\infty}^{+\infty} |D\mu| = \int_{-\infty}^{+\infty} d\mu^+(x) + \int_{-\infty}^{+\infty} d\mu^-(x) = 2$. Substituting in (3), with $f = 1$, we have the first equality of the thesis for $E(A)$. The second equality follows observing that $w$, as defined in Proposition 1, is equal to $1/2$ and using (6) in Proposition 1 with $f(\alpha) = 1$. □

### 3.3 Example

We now present an example to show that, when $f(\alpha) \neq 1$, we may have $w \neq 1/2$. We compute the mean values of two fuzzy numbers shown in Fig. 1 for $f(\alpha) = 2\alpha$. First, we consider the fuzzy number (a). We have $\int_{-\infty}^{+\infty} x\, f(\mu(x))\,d\mu^+(x) = 25/12$, $\int_{-\infty}^{+\infty} f(\mu(x))\,d\mu^+(x) = 5/4$ and thus $E_*(A; f) = 5/3$. Furthermore, $\int_{-\infty}^{+\infty} x\, f(\mu(x))\,d\mu^-(x) = 59/16$, $\int_{-\infty}^{+\infty} f(\mu(x))\,d\mu^-(x) = 17/16$ and then $E^*(A; f) = 59/17$. Moreover $E(A; f) = 277/111 \approx 2.5$ and $w = 17/37 \neq 1/2$.

We now consider the fuzzy number (b). We have $\int_{-\infty}^{+\infty} x\, f(\mu(x))\,d\mu^+(x) = 25/12$, $\int_{-\infty}^{+\infty} f(\mu(x))\,d\mu^+(x) = 5/4$ and thus $E_*(A; f) = 5/3$. Furthermore, $\int_{-\infty}^{+\infty} x\, f(\mu(x))\,d\mu^-(x) = 55/12$, $\int_{-\infty}^{+\infty} f(\mu(x))\,d\mu^-(x) = 5/4$ and then $E^*(A; f) = 11/3$. Moreover $E(A; f) = 8/3 \approx 2.7$ and $w = 1/2$.

### 3.4 Possibilistic Framework

We observe that our proposal, which starts from a variational measure, may be viewed in a possibilistic framework. Indeed, since $Poss(A \le a_L(\alpha)) = sup_{x \le a_L(\alpha)}\mu(x) = \mu(a_L(\alpha))$ and $Poss(A \ge a_R(\alpha)) = sup_{x \ge a_R(\alpha)}\mu(x) = \mu(a_R(\alpha))$, from (6) we get

$$E_*(A; f) = \int_0^1 a_L(\alpha)\, f(Poss(A \le a_L(\alpha)))\,d\alpha \Big/ \int_0^1 f(Poss(A \le a_L(\alpha)))\,d\alpha\,,$$

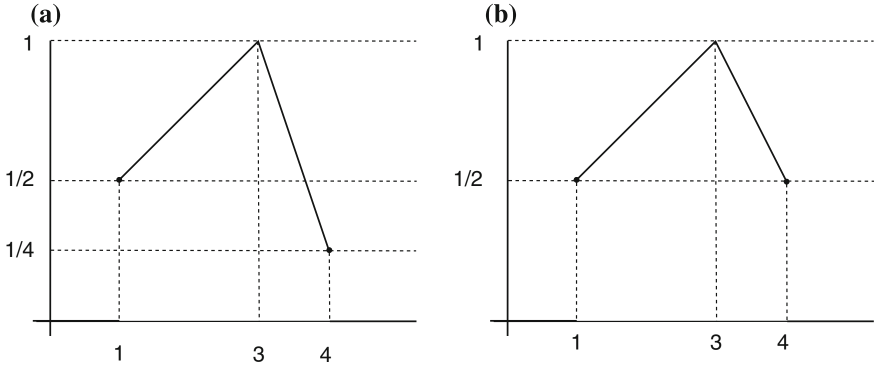


**Fig. 1** Two fuzzy numbers with non-continuous membership function

$$E^*(A;f) = \int_0^1 a_R(\alpha)\, f(Poss(A \geq a_R(\alpha)))\, d\alpha \Big/ \int_0^1 f(Poss(A \geq a_R(\alpha)))\, d\alpha.$$

Note that $A$ does not need to be continuous. In the special case when $A$ is a continuous fuzzy number we retrieve the lower and upper possibilistic mean values proposed in [3] and [5] (see Proposition 4). Thus our approach offers an extension of possibilistic mean values to the case of fuzzy numbers whose membership functions are upper-semicontinuous but are not necessarily continuous.

## 3.5 Weighted Mean Value of Continuous Fuzzy Numbers

If we use the variational approach for continuous fuzzy numbers we obtain that the weighted mean value is the simple average of upper and lower $f$-weighted mean values. This means that for every weighting function $f$ the weight $w = 1/2$.

**Proposition 4** *If $A$ is a fuzzy number with continuous membership function $\mu$ then $E(A;f) = \frac{1}{2}\int_{-\infty}^{+\infty} x\, f(\mu(x))\, |D\mu| = \frac{1}{2}\int_0^1 (a_L(\alpha) + a_R(\alpha))\, f(\alpha)\, d\alpha$, and $E_*(A;f) = \int_{-\infty}^{+\infty} x f(\mu(x))\, d\mu^+(x) = \int_0^1 a_L(\alpha) f(\alpha)\, d\alpha$, $E^*(A;f) = \int_{-\infty}^{+\infty} x f(\mu(x))\, d\mu^-(x) = \int_0^1 a_R(\alpha) f(\alpha)\, d\alpha$. Moreover, $w = 1/2$ and thus $E(A;f) = (E_*(A;f) + E^*(A;f))/2$.*

*Proof* The assertions easily follow from previous results observing that, since $\mu$ is continuous, we have $\mu(a_L(\alpha)) = \alpha$ and $\mu(a_R(\alpha)) = \alpha$. In particular, we have $\int_{-\infty}^{+\infty} f(\mu(x))\, d\mu^+(x) = \int_0^1 f(\mu(a_L(\alpha)))\, d\alpha = \int_0^1 f(\alpha)\, d\alpha = 1$ and, similarly, $\int_{-\infty}^{+\infty} f(\mu(x))\, d\mu^-(x) = 1$. □

# 4 Weighted Variance of Fuzzy Numbers

In this section we introduce two definitions of $f$-weighted variance of a fuzzy number $A$ whose membership function is upper semicontinuous. The first, $Var_1(A)$, is obtained as the simple average of the lower and upper variances of $A$, that derive from the definition of lower and upper mean value of $A$ given in Definition 2. The second, $Var_2(A)$, is the natural definition of variance that starts from Definition 1. In the continuous case and for particular weights $f$ we recover classical definitions introduced by other authors.

**Definition 3** We define the lower and upper variances of $a$, respectively, as

$$Var_*(A;f) = \int_{-\infty}^{+\infty} (x - E_*(A;f))^2 \; f(\mu(x))\, d\mu^+(x) \Big/ \int_{-\infty}^{+\infty} f(\mu(x))\, d\mu^+(x)\,,$$
$$Var^*(A;f) = \int_{-\infty}^{+\infty} \left(x - E^*(A;f)\right)^2 \; f(\mu(x))\, d\mu^-(x) \Big/ \int_{-\infty}^{+\infty} f(\mu(x))\, d\mu^-(x)\,.$$

and the variance of $A$ as $Var_1(A;f) = (Var_*(A;f) + Var^*(A;f))/2$.

Using (7) and (8) with $g(x) = (x - E_*(A;f))^2$ and $g(x) = (x - E^*(A;f))^2$, respectively, we may express previous definitions in terms of $\alpha$-cuts as follows

$$Var_*(A;f) = \int_0^1 (a_L(\alpha) - E_*(A;f))^2 \; f(\mu(a_L(\alpha)))\, d\alpha \Big/ \int_0^1 f(\mu(a_L(\alpha)))\, d\alpha,$$
$$Var^*(A;f) = \int_0^1 \left(a_R(\alpha) - E^*(A;f)\right)^2 \; f(\mu(a_R(\alpha)))\, d\alpha \Big/ \int_0^1 f(\mu(a_L(\alpha)))\, d\alpha.$$

**Definition 4** Alternatively, we may define the variance of a fuzzy number $A$ as $Var_2(A;f) = \int_{-\infty}^{+\infty} (x - E(A;f))^2 \, f(\mu(x))\, |D\mu| \Big/ \int_{-\infty}^{+\infty} f(\mu(x)) |D\mu|$.

## *4.1 Weighted Variance of Continuous Fuzzy Numbers*

**Proposition 5** *If $A$ is a fuzzy number with continuous membership function $\mu$ then $Var_*(A;f) = \int_0^1 (a_L(\alpha) - E_*(A;f))^2 \; f(\alpha)\, d\alpha$ and $Var^*(A;f) = \int_0^1 (a_R(\alpha) - E^*(A;f))^2 \; f(\alpha)\, d\alpha$.*

*Remark 1* If $A$ is a continuous fuzzy number and $f(\alpha) = 2\alpha$ we obtain that $Var_1(A)$ matches the variance proposed by Zhang and Nie [9].

**Proposition 6** *If $A$ is a continuous fuzzy number then*
*$Var_2(A;f) = \frac{1}{2} \int_0^1 \left((a_L(\alpha) - E(A;f))^2 + (a_R(\alpha) - E(A;f))^2\right) f(\alpha)\, d\alpha$.*

*Remark 2* If $A$ is a continuous fuzzy number and $f(\alpha) = 2\alpha$ we obtain that $Var_2(A; f)$ agrees with the variance $Var'(A)$ introduced by Carlsson and Fullér [3].

**Proposition 7** *If A is a fuzzy number with differentiable membership function* $\mu$ *then* $Var_2(A; f) = \frac{1}{2} \int_{-\infty}^{+\infty} (x - E(A; f))^2 f(\mu(x)) \, |\mu'(x)| \, dx$.

*Remark 3* If $A$ is a fuzzy number with differentiable membership function $\mu$ and $f(\alpha) = 2\alpha$ then $Var_2(A)$ is the variance proposed by Li, Guo and Yu [8].

## 5 Conclusion

In this paper we have presented the weighted mean value and variance definitions for fuzzy numbers with upper semicontinuous membership functions that are of bounded variation. In the same context in a previous paper we have looked for a definition of evaluation and ranking for fuzzy numbers [1]. The idea to use the space of bounded variation functions has allowed us to view the mean value and the variance of fuzzy numbers either in a variational context or in a classical way by $\alpha$-cuts. This choice, thanks to the freedom that we have left to the weights, has shown interesting results that are able to generalize the previous ones but even to recover others present in literature for the continuous case. We have intention to continue in this direction to see if the space of bounded variation functions offers a new research field for fuzzy sets.

## References

1. Anzilli L, Facchinetti G (2013) The total variation of bounded variation functions to evaluate and rank fuzzy quantities. Int J Intell Syst 28(10):927–956
2. Bede B (2013) Mathematics of fuzzy sets and fuzzy logic. Springer
3. Carlsson C, Fullér R (2001) On possibilistic mean value and variance of fuzzy numbers. Fuzzy Sets Syst 122(2):315–326
4. Dubois D, Prade H (1987) The mean value of a fuzzy number. Fuzzy Sets Syst 24(3):279–300
5. Fullér R, Majlender P (2003) On weighted possibilistic mean and variance of fuzzy numbers. Fuzzy Sets Syst 136(3):363–374
6. Georgescu I (2009) Possibilistic risk aversion. Fuzzy Sets Syst 160(18):2608–2619
7. Leoni G (2009) A first course in Sobolev spaces, vol 105. American Mathematical Society Providence, RI
8. Li X, Guo S, Yu L (2015) Skewness of fuzzy numbers and its applications in portfolio selection. Fuzzy Syst IEEE Trans 23(6):2135–2143
9. Zhang WG, Nie ZK (2003) On possibilistic variance of fuzzy numbers. In: Rough sets, fuzzy sets, data mining, and granular computing.Springer, pp 398–402

# On the Construction of Radially Symmetric Trivariate Copulas

**José De Jesús Arias García, Hans De Meyer and Bernard De Baets**

**Abstract** We propose a method to construct a 3-dimensional symmetric function that is radially symmetric, using two symmetric 2-copulas, with one of them being also radially symmetric. We study the properties of the presented construction in some specific cases and provide several examples for different families of copulas.

**Keywords** Copula · Quasi-copula · Radial symmetry · Aggregation function

## 1 Introduction

An $n$-dimensional copula (or, for short, $n$-copula) is a multivariate distribution function with the property that all its $n$ univariate marginals are uniform distributions on $[0, 1]$. Formally, an $n$-copula is a $[0, 1]^n \to [0, 1]$ function that satisfies the following conditions:

1. $C_n(\mathbf{x}) = 0$ if $\mathbf{x}$ is such that $x_j = 0$ for some $j \in \{1, 2, \ldots, n\}$.
2. $C_n(\mathbf{x}) = x_j$ if $\mathbf{x}$ is such that $x_i = 1$ for all $i \neq j$.
3. $C_n$ is $n$-increasing, i.e., for any $n$-box $\mathbf{P} = [a_1, b_1] \times \cdots \times [a_n, b_n] \subseteq [0, 1]^n$ it holds that
$$V_{C_n}(\mathbf{P}) = \sum_{\mathbf{x} \in \text{vertices}(\mathbf{P})} (-1)^{S(\mathbf{x})} C_n(\mathbf{x}) \geq 0\,,$$

J.D.J. Arias García (✉) · B. De Baets
KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, 9000 Ghent, Belgium
e-mail: JoseDeJesus.AriasGarcia@ugent.be

B. De Baets
e-mail: Bernard.DeBaets@ugent.be

H. De Meyer
Department of Applied Mathematics and Computer Science, Ghent University,
Krijgslaan 281 S9, 9000 Ghent, Belgium
e-mail: Hans.DeMeyer@ugent.be

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_2

where $S(\mathbf{x}) = \#\{j \in \{1, 2, \dots, n\} \mid x_j = a_j\}$. $V_{C_n}(P)$ is called the $C_n$-volume of $P$.

Due to Sklar's theorem, which states that any continuous multivariate distribution function can be written in terms of its $n$ univariate marginals by means of a unique $n$-copula, $n$-copulas have become one of the most important tools for the study of certain types of properties of random vectors, such as stochastic dependence (see [4, 13] for more details on $n$-copulas). One example of a property that can be directly studied from copulas is the property of radial symmetry. An $n$-dimensional random vector $(X_1, \dots, X_n)$ is said to be radially symmetric about $(x_1, \dots, x_n)$ if the distribution of the random vector $(X_1 - x_1, \dots, X_n - x_n)$ is the same as the distribution of the random vector $(x_1 - X_1, \dots, x_n - X_n)$. It is easily shown that radial symmetry can be characterized using copulas: a random vector $(X_1, \dots, X_n)$ is radially symmetric about $(x_1, \dots, x_n)$ if and only if for any $j \in \{1, \dots, n\}$ $X_j - x_j$ has the same distribution as $x_j - X_j$ and the $n$-copula $C_n$ associated to the random vector satisfies the identity $C_n = \hat{C}_n$, where $\hat{C}_n$ denotes the survival $n$-copula associated to $C_n$, and which can be computed as:

$$\begin{aligned}\hat{C}_n(x_1, \dots, x_n) = \sum_{j=1}^{n} x_j - (n-1) + \sum_{i<j}^{n} C_n(1, \dots, 1 - x_i, \dots, 1 - x_j, \dots, 1) \\ - \sum_{i<j<k}^{n} C_n(1, \dots, 1 - x_i, \dots, 1 - x_j, \dots 1 - x_k, \dots, 1) + \dots \\ +(-1)^n C_n(1 - x_1, 1 - x_2, \dots, 1 - x_n)\,. \qquad (1)\end{aligned}$$

Due to this characterization, we say that an $n$-copula $C_n$ is radially symmetric if it satisfies the identity $C_n = \hat{C}_n$. Survival copulas also have a probabilistic interpretation. If the random vector $(X_1, \dots X_n)$ has the copula $C_n$ as its distribution function, then $\hat{C}_n$ is the distribution function of the random vector $(1 - X_1, \dots, 1 - X_n)$. This probabilistic interpretation has led to several studies of the transformations of copulas which are induced by certain types of transformations on random variables (see [6–8]). These transformations have been generalized and studied in the framework of aggregation functions [1–3].

Radially symmetric copulas have a particular importance in stochastic simulation, as they are used as a part of the multivariate version of the antithetic variates method, which is a variance reduction technique used in Monte Carlo methods (see [12]). In the binary case, well-known examples of families of bivariate copulas that are radially symmetric are the Gaussian family, the Frank family and the Farlie-Gumbel-Morgenstern (FGM) family. However, there are few attempts to construct families of $n$-copulas with some specific properties for $n \geq 3$. In this contribution we propose a construct method for trivariate radially symmetric copulas.

## 2 Radial Symmetry and Associativity

Archimedean $n$(-quasi)-copulas are one of the most well-known classes of $n$-copulas, which have the additional property that they are symmetric, i.e., for any permutation $\sigma$ of $\{1, 2, \dots, n\}$ and for any $\mathbf{x} \in [0, 1]^n$ it holds that

$$C_n(x_1, x_2, \dots, x_n) = C_n(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}) .$$

Archimedean $n$(-quasi)-copulas are also associative and can be defined recursively, i.e., for any $x, y, z \in [0, 1]$ the equality $C_2(x, C_2(y, z)) = C_2(C_2(x, y), z)$ holds and for any $n \geq 2$ and $\mathbf{x} \in [0, 1]^{n+1}$ it holds that

$$\begin{aligned} C_{n+1}(\mathbf{x}) &= C_2(x_1, C_n(x_2, \dots, x_{n+1})) \\ &= C_2(C_n(x_1, \dots, x_n), x_{n+1}) . \end{aligned}$$

It is well known that Archimedean $n$(-quasi)-copulas can be fully characterized in terms of an additive generator (see [11]). Some further generalizations of Archimedean copulas have been proposed; for example in [10], nested Archimedean $n$-copulas are studied, where several bivariate Archimedean copulas are iterated to construct an $n$-copula (for example, in the trivariate case $C_2(x, D_2(y, z))$ would be an example of such a construction, where $C_2$ and $D_2$ are bivariate Archimedean copulas).

If a 2-copula is radially symmetric, then for any $(x, y) \in [0, 1]^2$, it holds that

$$C_2(x, y) + (1 - C_2(1 - x, 1 - y)) = x + y .$$

If $C_2$ is an Archimedean copula, the latter equation is a particular case of a functional equation studied by Frank in [5]. More specifically in [5], it is proven that if a continuous function $F : [0, 1]^2 \to [0, 1]$ satisfies the following properties

1. for any $x \in [0, 1]$ it holds that $F(x, 0) = F(0, x) = 0$,
2. for any $x \in [0, 1]$ it holds that $F(x, 1) = F(1, x) = x$,
3. the functions $F(x, y)$ and $G(x, y) = x + y - F(x, y)$ are both associative;

then $F$ must be a member of the Frank family of copulas or an ordinal sum constructed from members of this family. The Frank family is given by:

$$F^{(\alpha)}(x, y) = -\frac{1}{\alpha} \ln \left( 1 + \frac{(e^{-\alpha x} - 1)(e^{-\alpha y} - 1)}{e^{-\alpha} - 1} \right) ,$$

where $\alpha \in \mathbb{R} \cup \{-\infty, \infty\}$ (Although the case $\alpha = \infty$ is not an Archimedean copula).

In [8] this result is complemented by showing that 2-copulas that are both associative and radially symmetric are members of the Frank family of copulas or an ordinal sum of the form $C_2 = (\langle a_j, b_j, F_{(j)}^{(\alpha_j)} \rangle)_{j \in J}$, such that for any $j$, there exists $i_j$ with the property that $\alpha_j = \alpha_{i_j}$, $a_j = 1 - b_{i_j}$ and $b_j = 1 - a_{i_j}$.

Note that if a 3-copula is radially symmetric, then its 2-dimensional marginals must also be radially symmetric. This trivially generalizes to higher dimensions. Hence if an associative 3-copula is radially symmetric, then its 2-dimensional marginals must also be solutions of the Frank functional equation. Unfortunately, as shown in [9], for $n \geq 3$ the only solutions are the product copula $\Pi_n(x_1, \ldots, x_n) = x_1 x_2 \ldots x_n$ (which is the copula of independent random variables) and the minimum operator $M_n(x_1, \ldots, x_n) = \min(x_1, \ldots, x_n)$ (which is the copula of comonotonic random variables) or ordinal sums constructed using these two $n$-copulas. From this it follows that if we want to construct radially symmetric copulas in higher dimensions, we must weaken the condition of associativity (and as a consequence the Archimedean property is lost), as jointly requiring both properties is too restrictive in higher dimensions.

## 3 The Construction

In [8], the authors study several transformations of bivariate copulas. They show that every radially symmetric 2-copula has the following form

$$\frac{C_2(x, y) + \hat{C}_2(x, y)}{2}, \tag{2}$$

where $C_2$ is a 2-copula. This result is easily generalized to higher dimensions. Keeping this result in mind, we propose the following construction method in three dimensions.

**Definition 1** Let $C_2, D_2$ be two symmetric 2-copulas, such that $C_2$ is also radially symmetric. We define the symmetric function $S_{C_2,D_2} : [0, 1]^3 \to \mathbb{R}$ associated to the pair $C_2, D_2$ as

$$\begin{aligned} S_{C_2,D_2}(x, y, z) = &\frac{1}{2}[1 - x - y - z + C_2(x, y) + C_2(x, z) + C_2(y, z)] \\ &+\frac{1}{2}[H_{C_2,D_2}(x, y, z) - H_{C_2,D_2}(1 - x, 1 - y, 1 - z)], \end{aligned} \tag{3}$$

where

$$\begin{aligned} H_{C_2,D_2}(x, y, z) = &D_2(x, C_2(y, z)) + D_2(y, C_2(x, z)) + D_2(z, C_2(x, y)) \\ &-\frac{2}{3}[D_2(x, D_2(y, z)) + D_2(y, D_2(x, z)) + D_2(z, D_2(x, y))]. \end{aligned}$$

Note that the function $H_{C_2,D_2}$ satisfies the boundary conditions of a 3-copula, and from this is easy to prove that $S_{C_2,D_2}$ also satisfies the boundary conditions of a copula, and that the bivariate marginals of $S_{C_2,D_2}$ are all equal to $C_2$.

**Proposition 1** *Let $C_2$, $D_2$ be two symmetric 2-copulas, such that $C_2$ is also radially symmetric. Let $S_{C_2,D_2}$ be defined as in Eq. (3). If $S_{C_2,D_2}$ is a 3-copula, then it is radially symmetric.*

*Proof* It can be shown after some tedious computations that we can rewrite $S_{C_2,D_2}$ as

$$\begin{aligned} S_{C_2,D_2}(x,y,z) = x + y + z - 2 + S_{C_2,D_2}(1-x,1-y,1) \\ +S_{C_2,D_2}(1-x,1,1-z) + S_{C_2,D_2}(1,1-y,1-z) \\ -S_{C_2,D_2}(1-x,1-y,1-z)\,, \end{aligned}$$

i.e., from the definition of survival copula given by Eq. (1), it follows that if $S_{C_2,D_2}$ is a 3-copula, then $S_{C_2,D_2}$ is a radially symmetric 3-copula because it coincides with its associated survival copula. □

However, $S_{C_2,D_2}$ is not necessarily a 3-copula, as it may even not be an increasing function. For example, if $C_2 = D_2 = F^{(-2)}$, we can easily see that $S_{F^{(-2)},F^{(-2)}}$ $(\frac{1}{2}, \frac{1}{10}, \frac{1}{10}) < 0 = S_{F^{(-2)},F^{(-2)}}(0, \frac{1}{10}, \frac{1}{10})$. We now provide some examples where the construction effectively yields a 3-copula.

*Example 1* Consider the Frank family of 2-copulas. From the results in [11], we know that for $\alpha \geq -\ln(2)$, the 3-dimensional version of the Frank 2-copula, given by $F_3^{(\alpha)}(x,y,z) = F^{(\alpha)}(x, F^{(\alpha)}(y,z))$, is a 3-copula. From this, it follows immediately that if $\alpha \geq -\ln(2)$ then $S_{F^{(\alpha)},F^{(\alpha)}}$ is a 3-copula. However, with some computational help, it can be shown that $S_{F^{(\alpha)},F^{(\alpha)}}$ is also a 3-copula for $\alpha \geq -\ln(3)$.

*Example 2* The FGM family of bivariate copulas is given by

$$F^{(\theta)}(x,y) = xy + \theta xy(1-x)(1-y)\,, \quad \theta \in [-1,1]\,.$$

In this case, some computations show that $S_{F^{(\theta)},F^{(\theta)}}$ is a 3-copula if and only if $\theta \in [-1/2(3-\sqrt{5}), 1/2(\sqrt{21}-3)]$.

*Example 3* For any 2-copula $C_2$, $S_{C_2,\Pi_2}$ is a 3-copula if and only if for any $x_1, x_2, y_1, y_2, z_1, z_2 \in [0,1]$ such that $x_1 \leq x_2$, $y_1 \leq y_2$, $z_1 \leq z_2$ it holds that

$$\begin{aligned} &(x_2 - x_1)V_{C_2}([y_1,y_2]\times[z_1,z_2]) + (y_2 - y_1)V_{C_2}([x_1,x_2]\times[z_1,z_2]) \\ &+(z_2 - z_1)V_{C_2}([x_1,x_2]\times[y_1,y_2]) \\ &\geq 2(x_2-x_1)(y_2-y_1)(z_2-z_1)\,. \end{aligned}$$

If $C_2$ is absolutely continuous, then this last condition is equivalent to:

$$\frac{\partial^2 C_2}{\partial x \partial y}(x,y) + \frac{\partial^2 C_2}{\partial x \partial z}(x,z) + \frac{\partial^2 C_2}{\partial y \partial z}(y,z) \geq 2\,. \tag{4}$$

for almost every $x, y, z \in [0, 1]$. An example of a family of 2-copulas that satisfies Eq. (4) is the FGM copulas $F^{(\theta)}$ for $\theta \in [-1/3, 1/3]$.

We note that Example 3 can be generalized to higher dimensions. Given a symmetric $n$-copula $C_n$ that satisfies $C_n = \hat{C}_n$, we define the $(n+1)$-dimensional function $S_{C_n}$ as

$$\begin{aligned} S_{C_n}(x_1, \ldots, x_{n+1}) = \frac{1}{2}\Big[ &\sum_{j=1}^{n+1} x_j - n + \sum_{i<j}^{n+1} C_n(1, \ldots, 1-x_i, \ldots, 1-x_j, \ldots, 1) \\ &- \sum_{i<j<k}^{n+1} C_n(1, \ldots, 1-x_i, \ldots, 1-x_j, \ldots 1-x_k, \ldots, 1) + \ldots \\ &+ (-1)^n \sum_{j=1}^{n+1} C_n(1-x_1, \ldots, 1-x_{j-1}, 1-x_{j+1} 1-x_{n+1})\Big] \\ &+ \frac{1}{2}\left[H(x_1, \ldots, x_{n+1}) + (-1)^{n+1} H(1-x_1, \ldots, 1-x_{n+1})\right], \end{aligned}$$

where

$$\begin{aligned} H(x_1, \ldots, x_{n+1}) = &\sum_{j=1}^{n+1} x_j C_n(\mathbf{x}_{\{j\}}) \\ &- \sum_{i<j}^{n+1} x_i x_j C_{n-1}(\mathbf{x}_{\{i,j\}}) \\ &\ldots \\ &+ (-1)^n \sum_{i<j}^{n+1} \left( \prod_{k \neq i,j} x_k \right) C_2(x_i, x_j) \\ &+ n(-1)^{n+1} x_1 x_2 \ldots x_{n+1}, \end{aligned}$$

and $\mathbf{x}_A$ denotes the vector whose components take the values of the elements $x_1, \ldots, x_{n+1}$, except of those elements $x_j$ for which $j$ is in the set $A$ of indices. It can be proven that the function $S_{C_n}$ is such that if $S_{C_n}$ is an $(n+1)$-copula, then it is an $(n+1)$-dimensional radially symmetric copula, with $n$-dimensional marginals given by $C_n$. The characterization in the absolutely continuous case is also simple, since after doing some combinatorial analysis, it is easy to prove that if $C_n$ is absolutely continuous, then $S_{C_n}$ is an $(n+1)$-copula if and only if for any $x_1, \ldots x_{n+1} \in [0, 1]$, it holds that

$$\sum_{j=1}^{n+1} \frac{\partial^n C_n}{\partial x_1 \ldots \partial x_{j-1} \partial x_{j+1} \ldots \partial x_{n+1}} (\mathbf{x}_{\{j\}})$$

$$
\begin{aligned}
& -\sum_{i<j}^{n+1} \frac{\partial^{n-1} C_{n-1}}{\partial x_1 \ldots \partial x_{i-1} \partial x_{i+1}, \ldots, \partial x_{j-1} \partial x_{j+1}, \ldots \partial x_{n+1}} (\mathbf{x}_{\{i,j\}}) \\
& \ldots \\
& +(-1)^n \sum_{i<j}^{n+1} \frac{\partial^2 C_2}{\partial x_i \partial x_j}(x_i, x_j) + n(-1)^{n+1} \geq 0 \,.
\end{aligned}
$$

## 4 Conclusions and Future Work

We proposed a way of constructing a symmetric and radially symmetric trivariate copula with given bivariate marginals, and provided some examples of this construction. However, it remains is an open problem to determine for which pairs of 2-copulas $C_2$ and $D_2$, the function $S_{C_2,D_2}$ is a 3-copula. A first step in this study would be to characterize the set of ternary aggregation functions for which it holds that their 'survival transformation' is also an aggregation function (see [1–3]). A final task is to analyse whether the presented construction can be properly generalized to any dimension $n > 3$.

**Acknowledgments** The first author was supported by the "Consejo Nacional de Ciencia y Tecnología" (México) grant number 382963.

## References

1. De Baets B, De Meyer H, Kalická J, Mesiar R (2009) Flipping and cyclic shifting of binary aggregation functions. Fuzzy Sets Syst 160(6):752–765
2. De Baets B, De Meyer H, Mesiar R (2012) Binary survival aggregation functions. Fuzzy Sets Syst 191:83–102
3. Durante F, Fernández-Sánchez J, Quesada-Molina JJ (2014) Flipping of multivariate aggregation functions. Fuzzy Sets Syst 252:66–75
4. Durante F, Sempi C (2015) Principles of Copula Theory. CRC Press
5. Frank MJ (1979) On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. Aequationes Mathematicae 19(1):194–226
6. Fuchs S (2014) Multivariate copulas: Transformations, symmetry, order and measures of concordance. Kybernetika 50(5):725–743
7. Fuchs S, Schmidt KD (2014) Bivariate copulas: transformations, asymmetry and measures of concordance. Kybernetika 50(1):109–125
8. Klement EP, Mesiar R, Pap E (2002) Invariant copulas. Kybernetika 38(3):275–286
9. Kuková M, Navara M (2013) Principles of inclusion and exclusion for fuzzy sets. Fuzzy Sets Syst 232:98–109
10. McNeil AJ (2008) Sampling nested Archimedean copulas. J Stat Comput Simul 78(6):567–581
11. McNeil AJ, Nešlehová J (2009) Multivariate Archimedean copulas, d-Monotone functions and $l_1$-norm symmetric distributions, Ann Stat 3059–3097
12. Mai JF, Scherer M (2012) Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications, vol 4. World Scientific
13. Nelsen R (2006) An Introduction to Copulas. Springer, New York

# Simulation of the Night Shift Solid Waste Collection System of Phuket Municipality

**Somsri Banditvilai and Mantira Niraso**

**Abstract** This research was conducted in order to simulate the night shift solid waste collection system of Phuket Municipality, Thailand. The Phuket Municipality faced the problems of residual waste and an unbalanced load for solid waste collection teams. The waste management committee of Phuket Municipality wanted to improve the solid waste collection system to run more efficiently. This research analyzed the volume of solid waste collection instead of the weight, and has separated the solid waste collection points into 11 "types". The data was collected from the survey form. Minitab 16.1 was used to analyze and test the data distribution, and then used them to build the model. Microsoft Visual C++ was used to build the simulation model, which was then verified and validated extensively. The model represented the actual night shift solid waste collection system of Phuket Municipality. The heuristic approach was then employed to apply new assigned zones and routings. The results from the study of the new system of night shift solid waste collection system of Phuket Municipality shows that there is no residual waste and no unbalanced load between solid waste collection teams. The new system works effectively and can decrease the total number of trips for solid waste collection by 9.1 % and the average distance and time for the solid waste collection system are decreased by 7.42 % and 7.10 % respectively.

## 1 Introduction

Phuket is a popular tourist destination in Thailand. There are historical landmarks and it has a beautiful natural scenery. The number of tourists, both domestic and foreign, is increasing exponentially every year. As a result, there is a rapid expansion of establishments such as hotels, restaurants, hospitals etc. These establishments also

S. Banditvilai (✉)
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
e-mail: somsri_b2000@yahoo.com

M. Niraso (✉)
Institute of Physical Education Trang, Trang, Thailand
e-mail: meniraso@hotmail.com

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_3

pollute the environment and increase solid waste, especially in the Phuket Municipality. Solid waste management starts with collection, which is followed by transportation, processing and disposal or destruction. The collection and transportation of solid waste are considered very important, and accounts for 75 to 80 % of the total cost [3]. Therefore, this research is focused on the collection and transportation. It is essential to manage solid waste collection efficiency, to save time and costs. It is difficult to design a solid waste collection system to suit the garbage problems that is always changing. Simulation is the imitation of the operation of a real-world process or system over time [2]. In order to increase the vehicle productivity, it is necessary to plan vehicle routing so that the quantity of waste collected is maximized [6]. Therefore, a simulation model is a suitable tool in designing of the solid waste collection system. This research used the simulation model in assigning areas of responsibilities and routes for solid waste collection trucks.

## 2 The Scope of This Research

The Phuket Municipality is responsible for solid waste collection of 14 km$^2$. This study covers the night shift solid waste collection system routing and area of responsibility of each trucks, and not covered by day shift and private solid waste collection companies. The traffic conditions are not considered, since it is the night shift and there are no traffic jams.

## 3 Methodology

[1, 5, 6] and others studied the solid waste collection by analyzing the weight of solid waste collection. However, this research analyzed the volume of solid waste collection instead of the weight, since a full truck means that the solid waste has reached the limit of truck volume not the weight limit. The volume of solid waste can easily determine by the container size. The heuristic approach includes the identification of preliminary routes to be updated to balance routes through trial and error [4]. Therefore, this research employed the heuristic approach in assigning zones and routes. Since the volume of solid waste collection and time spent at each collection point are depending on the place that we collected the garbage is called "collection point". This research has classified the collection points into 11 types which are commercial buildings, restaurants, hotels, markets, government buildings, gas stations and garages, religious places, schools, hospitals, residential areas, streets/parks.

### *3.1 The Study of the Night Solid Waste Collection System of Phuket Municipality*

Currently Phuket Municipality separates the area of night shift solid waste collection system into 10 zones with one truck being responsible for each zone as shown in Fig. 1. There are 11 compaction trucks-10 currently used and 1 spare truck. The loading capacities of the trucks range from 8 to 11 m$^3$. The Head of Cleaning Division is responsible for assigning the area of responsibility for each truck which is called a zone. The driver decides the route for solid waste collection for his zone. The night shift solid waste collection is daily performed on each street of the city and it goes from 9.00 p.m. to 5.00 a.m. The team of solid waste collection is composed of one driver and two garbage collectors. Each day the team will continue to operate until all the solid waste collection points are visited or the operation hours have ended. The collection trip will start from the station of Phuket Municipality. Then it runs from one solid waste collection point to the next in the area of responsibility until it has reached full capacity of truck or runs through all collection points. This, leads to the transfer of the solid waste to the disposal site, and the trip ends. Then starts another trip until all collection points are visited or the operation hours end at 5.00 a.m. After transferring solid waste to the disposal site, the driver drives the truck to the station. Upon completion of the operation, it is typically 2–3 trips per day per truck.

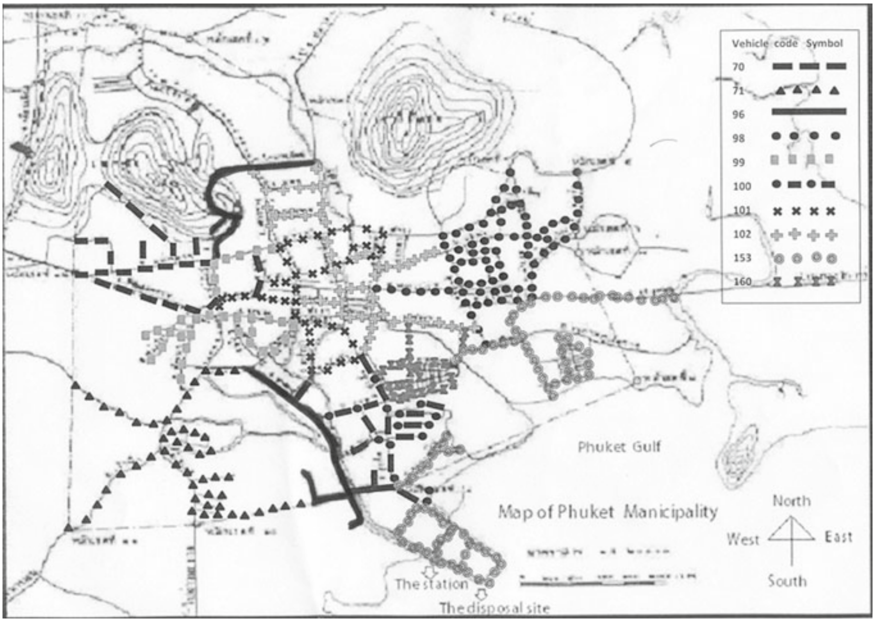


**Fig. 1** Phuket Municipal map before routing

### 3.2 Data Collection and Analysis

This research study has collected the general data of the solid waste collection system such as zoning areas, the number of solid waste collection vehicles, types, capacities of the vehicles, and the solid waste collection routes from the Cleaning Division of Public Health and the Environment Phuket Municipality, and maps of transportation networks from the Bureau of Engineering Phuket Municipality. The survey form designed for the operational solid waste collection system was used to record the operational data of trucks each day. Observations were made while riding in the collection vehicles. The detailed activities of the solid waste collection system were noted starting when the truck leaves the station until the operations of solid waste collection were completed, the collected data include: the transport route of each truck; the distance and time from the station to the first collection point; the volume of solid waste collection, and the time spent at the collection point classified by types of solid waste collection points; the distance and time between each collection point; the volume of solid waste transported in each trip by each vehicle; the distance and time from the final solid waste collection point to the disposal site; the transfer time at the disposal site; the distance and time from the disposal site to the station; the total distance and time of operations of each truck per day; the number of trips of each truck per day.

### 3.3 The Analysis of the Distribution

Minitab 16.1 was employed in analyzing the distribution of the collected data. Anderson-Darling test was used to test the distribution. It was found as follows: the number of solid waste collection points per vehicle code as shown in Table 1; the distance and time from the station to the first collection point of each vehicle code have a normal distribution with mean and variance as shown in Table 1; the volume of solid waste collection, and the time spent at the collection point classified by types of solid waste collection point have a normal distribution with mean and variance as shown in Table 2; the distance between each collection point has a normal distribution with a mean of 37.67 m and a variance of 31.26 $\text{m}^2$; the time to travel between the waste collection points has a normal distribution with a mean of 0.51 min and a variance of 0.24 $\text{min}^2$; the volume of solid waste that is fully capable of carrying by each type of vehicle have a normal distribution with mean and variance as shown in Table 3; the distance from the last collection point to the disposal site has a normal distribution with a mean of 4478 m and a variance of 3.06 $\text{m}^2$; the time to travel from the last collection point to the disposal site. From the survey data found that it has a normal distribution with mean 2.62 min and variance 1.69 $\text{min}^2$ The time spent at the disposal site has a normal distribution with a mean of 19.31 min and a variance of 8.65 $\text{min}^2$; the distance from the disposal site to the station is 1000 m; the time to travel from the disposal site to the station has a normal distribution with a mean of

**Table 1** The number of solid waste collection points, the distribution of distance (meters) and time (minutes) from the station to the first collection point of each vehicle code

| Vehicle code | Solid waste collection points | Distance from station to the first collection point | Time from station to the first collection point |
|---|---|---|---|
| 70 | 195 | N(2412.5, 1124.33) | N(7.69, 2.50) |
| 71 | 159 | N(2072.5, 66.12) | N(9.06, 2.37) |
| 96 | 84 | N(3312.5, 2140.39) | N(12.43, 6.81) |
| 98 | 125 | N(2125.0, 992.47) | N(8.94, 2.01) |
| 99 | 108 | N(2687.5, 210.02) | N(10.95, 4.45) |
| 100 | 142 | N(1237.5, 206.59) | N(3.72, 1.05) |
| 101 | 153 | N(1962.5, 998.48) | N(6.63, 2.92) |
| 102 | 120 | N(2712.5, 339.91) | N(10.44, 1.59) |
| 153 | 197 | N(2725.0, 237.55) | N(13.89, 6.42) |
| 160 | 104 | N(1925.0, 895.62) | N(6.81, 1.96) |

**Table 2** The distribution of the volume of solid waste collection (liters), and the time spent at the collection point (minutes) classified by types of solid waste collection point

| Types of solid waste collection point | The volume of solid waste collection | Time spent at the collection point |
|---|---|---|
| Commercial buildings | N(1218.57, 936.94) | N(8.25, 5.92) |
| Restaurants | N(266.67, 197.52) | N(2.68, 2.30) |
| Hotels | N(1173.18, 823.33) | N(9.94, 11.25) |
| Markets | N(2035.63, 2804.46) | N(12.25, 16.02) |
| Government buildings | N(644.17, 356.28) | N(2.40, 1.78) |
| Gas stations and garages | N(280.67, 132.10) | N(3.33, 2.39) |
| Religious places | N(441.25, 352.92) | N(3.0, 2.45) |
| Schools | N(1471.25, 890.47) | N(7.66, 6.40) |
| Hospitals | N(4625.00, 1957.13) | N(36.83, 17.71) |
| Residential areas | N(798.49, 460.01) | N(4.60, 3.10) |
| Streets/Parks | N(186.52, 113.0) | N(1.61, 1.09) |

**Table 3** The distribution of the volume of solid waste collection (liters) that is fully capable of being carried by each type of vehicle

| The compaction truck capacity | The volume of solid waste that is fully capable of carrying collection |
|---|---|
| 8000 | N(20743.93, 3714.06) |
| 10000 | N(24703.75, 2492.10) |
| 11000 | N(26148.57, 2191.21) |

4.15 min and a variance of 0.78 $\text{min}^2$; the number of trips of solid waste collection are 2 trips per day except only the vehicle code 101 and 102 that are 3 trips per day.

## 3.4 Model Building, Verification and Validation

By Microsoft Visual C++, a discrete-event simulation model of the night shift solid waste collection of Phuket Municipality was built. The model was verified and validated extensively in order to confirm that the model represents the current night shift solid waste collection system of Phuket Municipality. In a total of 1000 simulation runs per vehicle code, the distance, time and volume of solid waste collection between the current night shift solid waste collection system and the simulation model was compared as shown in Tables 4, 5 and 6.

**Table 4** The comparison of the average distance (kilometers) of solid waste collection between the current night shift solid waste collection system (actual) and the simulation model

| Vehicle code | Actual distance | Simulation distance | Difference(%) |
|---|---|---|---|
| 70 | 22.50 | 21.94 | 2.49 |
| 71 | 28.30 | 28.12 | 0.64 |
| 96 | 39.00 | 39.72 | −1.85 |
| 98 | 38.70 | 39.58 | −2.27 |
| 99 | 14.50 | 15.21 | −4.90 |
| 100 | 15.80 | 15.56 | 1.52 |
| 101 | 15.30 | 14.59 | 4.64 |
| 102 | 20.50 | 19.92 | 2.83 |
| 153 | 21.20 | 22.20 | −4.72 |
| 160 | 13.90 | 13.32 | 4.17 |

**Table 5** The comparison of the average time (hours) of solid waste collection between the current night shift solid waste collection system (actual) and the simulation model

| Vehicle code | Actual time | Simulation time | Difference(%) |
|---|---|---|---|
| 70 | 7.50 | 7.20 | 4.00 |
| 71 | 6.95 | 7.15 | −2.88 |
| 96 | 8.00 | 7.65 | 4.38 |
| 98 | 7.97 | 8.14 | −2.13 |
| 99 | 6.61 | 6.34 | 4.08 |
| 100 | 6.80 | 6.86 | −0.88 |
| 101 | 6.51 | 6.25 | 3.99 |
| 102 | 8.05 | 8.10 | −0.62 |
| 153 | 7.75 | 7.81 | −0.77 |
| 160 | 5.78 | 5.72 | 1.04 |

**Table 6** The comparison of the average volume (liters) of solid waste collection between the current night shift solid waste collection system (actual) and the simulation model

| Vehicle code | Actual volume | Simulation volume | Difference(%) |
|---|---|---|---|
| 70 | 32750 | 31159 | 4.86 |
| 71 | 37520 | 38949 | −3.81 |
| 96 | 34510 | 36014 | −4.36 |
| 98 | 48050 | 46256 | 3.73 |
| 99 | 38480 | 39294 | −2.12 |
| 100 | 32800 | 34214 | −4.31 |
| 101 | 47690 | 49708 | −4.23 |
| 102 | 45020 | 44509 | 1.14 |
| 153 | 38680 | 37067 | 4.17 |
| 160 | 38470 | 39510 | −2.70 |

Tables 4, 5 and 6 show that the difference between the average distance, time, and volume of the current night shift solid waste collection system and the simulation model are less than 5 %. The trips of each vehicle code are the same. Therefore, the simulation model represents the current night shift solid waste collection system of Phuket Municipality. Then the model can be used to evaluate alternative system to improve the solid waste collection system.

### 3.5 *Set the New Assigned Zones and Routings*

From the study of the night shift solid waste collection system of Phuket Municipality, the average operation time of each truck is between 5.72 and 8.14 h. The average distance of each truck is between 13.90 and 39.00 km. The trip of each vehicle is 2–3 trips per day. By employing the heuristic approach, the new assigned zones and routes was set to improve the night shift solid waste collection system in order to have the number of trips, distance and time of operation for each truck decreased. The Phuket Municipality has no residual waste, and each truck has similar hours of operation, regardless of the increase or decrease in the use of trucks, equipment, employees, and collection points.

## 4 Results and Conclusion

The results from the study of the new system of night shift solid waste collection system of Phuket Municipality shows that there is no residual waste and no unbalanced load between the teams of solid waste collection. The new assigned zones

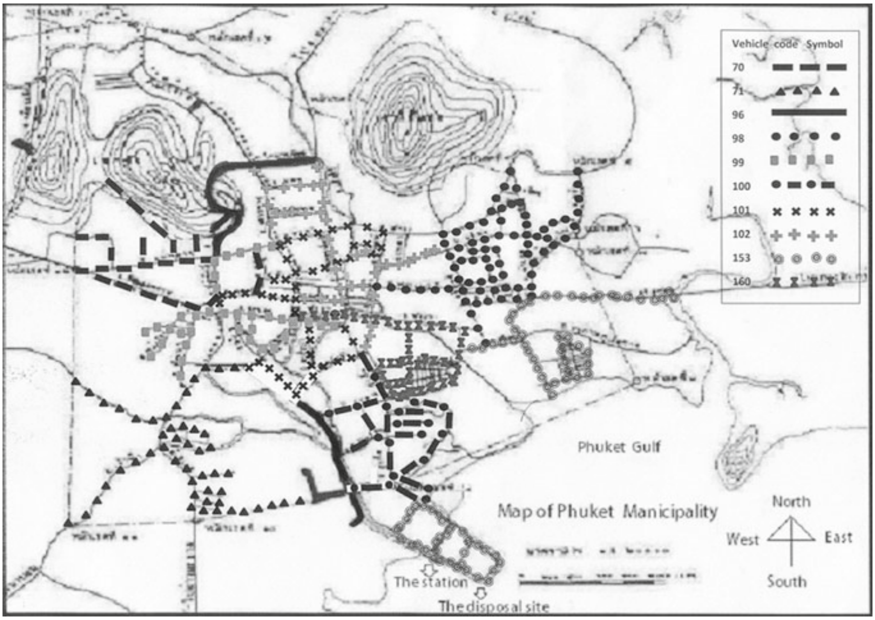


**Fig. 2** Phuket Municipal map after routing

(Fig. 2) and routes work effectively and can decrease the total number of trips for solid waste collection by 2 trips per day or 9.1 % and the average distance and time for the solid waste collection system are decreased by 7.42 % and 7.10 % respectively. The team of solid waste collection has the working hours between 6.26 and 7.23 h.

## References

1. Antmann ED et al (2012) Simulation-based optimization of solid waste management and recycling programs. In: lim G, Herrmann JW (eds) Proceeding of the 2012 Industrial and Systems Engineering Research Conference
2. Bank J, Carson JS, Nelson BL (2005) Discrete-Event System Simulation. Pearson Education, Upper Saddle River, NJ
3. Bhat VN (1996) A model for the optimal allocation of trucks for the solid waste management. Waste Manag Res 14:87–96
4. Bhide AD, Sundaresan BB (1983) Solid Waste Management in Developing Countries. Indian National Scientific Documentation Center, New Delhi
5. de Simonetto E, Borenstein D (2007) A decision support system for the operational planning of solid waste collection. Waste Manag 27:1286–1297
6. Tin AM et al (1995) Cost-benefit analysis of the municipal solid waste collection system in Yongon, Myanmar. Resour Conserv Recycl 14:103–131

# Updating Context in the Equation: An Experimental Argument with Eye Tracking

**Jean Baratgin, Brian Ocak, Hamid Bessaa and Jean-Louis Stilgenbauer**

**Abstract** The *Bayesian model* was recently proposed as a normative reference for psychology studies in deductive reasoning. This *new paradigm* supports that individuals evaluate the probability of an indicative conditional *if A then C* in the natural language as the conditional probability $P(C \textit{ given } A)$ ($P(C|A)$ according to *Bayes' rule*). In this paper, we show applying an *eye-tracking* methodology that if the cognitive process for both probability assessments ($P(\textit{if } A \textit{ then } C)$ and $P(C|A)$) is really identical, it actually doesn't match the traditional *focusing* situation of revision corresponding to Bayes' rule (change of reference class in a static universe). Individuals appear to revise their probability as if the universe was evolving. They use a *minimal rule* in mentally removing the elements of the worlds that are *not A*. This situation, called *updating*, actually seems to be the natural frame for individuals to evaluate the probability of indicative conditional and the conditional probability.

**Keywords** Equation · Conditional probability · Focusing · Updating · Eye-tracking methodology

## 1 New Paradigm in Psychology of Reasoning

For a decade, psychologists have argued that the binary logic was inadequate to account for the performance in reasoning tasks because people use strategies to reason under uncertainty (whose nature is probabilistic) [6, 23, 35, 38]. These authors

J. Baratgin (✉) · B. Ocak · H. Bessaa · J.-L. Stilgenbauer
Laboratoire CHArt, Université Paris 8 & EPHE, Site Paris-EPHE,
4-14 Rue Ferrus, 75014 Paris, France
e-mail: jean.baratgin@univ-paris8.fr

J. Baratgin
Institut Jean Nicod (École Normale Supérieure), Paris, France

B. Ocak
Université de Franche-Comté, Besançon, France

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_4

proposed to adopt the probabilistic *Bayesian theory*[1] [7]: Individuals' degrees of belief must respect the axioms of additive probability (static coherence) and their revision must follow the *conditioning principle* (dynamic coherence) which assumes that the revised probability of hypothesis $H$ upon learning the data $D$ ($P_D(H)$) at time $t_1$ is equal to the probability of $H$ conditioned on the (imagined or assumed) $D$ at time $t_0$ ($P(H|D)$ yielded by Bayes' rule[2]):

$$P_D(H) = P(H|D) \tag{1}$$

This shift of model of reference implies several conceptual and methodological modifications in the study of deductive reasoning [3, 4]. The main change is that the *indicative conditional 'If A (antecedent) then C (consequent)'* in natural language is interpreted as a single statement of a link (maybe weak) between $A$ and $C$ that does not necessarily coincide with the material conditional in formal logic (which is logically equivalent to $\neg A \vee C$).[3]

## 2 The Equation

De Finetti [17] distinguishes two levels of knowledge (belief and degree of belief) about the outcome of an event.

- For the belief, all events are conditional. Let $A$ be non-contradictory, the conditional event $C|A$ is *True* when $A \wedge C$ is true, *False* when $A \wedge \neg C$ is true and *Void* when $\neg A$ is true.[4] The fundamental relation for $C|A$ at this level is[5]:

[1] In this paper we refer to *de Finetti's subjective Bayesian probability* theory [16] that provides a unified perspective to study reasoning and probability judgment [10].

[2] The two usual forms of Bayes' rule are the *conditional probability*:

$$P(H|D) = \frac{P(H \wedge D)}{P(D)} = \frac{P(H \wedge D)}{P(H \wedge D) + P(\neg H \wedge D)}$$

and the *Bayes' identity*:

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} = \frac{P(H)P(D|H)}{P(H)P(D|H) + P(\neg H)P(D|\neg H)}$$

with $P(H)$ the prior probability of hypothesis $H$, $P(H|D)$ the posterior probability after the knowledge of data $D$, $P(D|H)$ the likelihood and $P(D)$ the probability of $D$.

[3] The other main change, not covered in this paper, is the analyzis of *deductive arguments* in the light of de Finetti's Bayesian coherence interval [12, 41, 45]. Some studies show a relative coherence in Human deduction under uncertainty [14, 41–43, 46, 48].

[4] Recent studies show that a majority of individuals have a trivalent interpretation of the conditional event [47] and that de Finetti's three-valued tables [18] are the best approximation for participants' truth tables [5, 6, 11].

[5] With $\wedge_k$ for the Kleene-Łukasiewicz-Heiting conjunction.

$$\textit{if } A \textit{ then } C = C|A = (C \wedge_k A)|A \tag{2}$$

- For the degree of belief, it is assumed that individuals are able to specify the *Void* value in subjective probabilistic terms [28, 39]. The main property (called *the Equation* [26]) is that the probability of an indicative conditional *if A then C* is equal to the probability of $C$ after the knowledge of $A$ (Ramsey test). It corresponds (by the conditioning principle) to the conditional probability and reads with Bayes' rule:

$$P(\textit{if } A \textit{ then } C) = P_A(C) = P(C|A) = \frac{P(C \wedge A)}{P(A)} \tag{3}$$

The majority of participants seem to act according to the Equation [13, 20, 21, 24, 25, 27, 36, 37, 40, 47, 48].

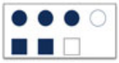
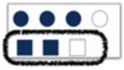
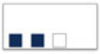
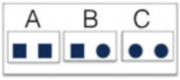
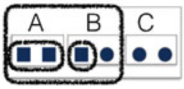
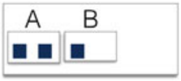
## 3 Context of Revision

Bayes' rule illustrates a singular situation of 'revision' called *focusing* [15, 22, 50]. It is assumed that one object is selected from the reference class and that a message releases information about this object. A change of reference class is consequently considered by *focusing* attention on a given subset of the initial reference class that complies with the information on the selected object. It concerns an atemporal revision in a *stable universe*; that is to say, a situation where the class of reference is not modified by the message. People seem to be 'good focusers' when the message concerns only one level of belief (indication on an object randomly extracted from a certain urn whose precise composition is known as the *chip problem* in Table 1, type 1). By contrast, very few participants grasp the focusing process when the message concerns two levels of belief (indication on an object extracted in two steps as *Bertrand's three boxes problem* in Table 1, type 2). Participants (and also experts) confronted to type 2 problems seem to naturally interpret (for pragmatic and cognitive reasons) [1, 2, 8, 9] the focusing situation as an *updating* [29, 32, 49, 50] situation of revision in which the reference class is evolving and the message conveys some information on this transformation (temporal revision). The new distribution of probability is obtained in two steps (*minimal rule*[6] [50]):

1. Representing the new class of reference that results from the modification of the initial one (removing the worlds invalidated by the message).
2. Inferring the 'new' prior probability distribution.

[6] This minimal rule is actually the intuitive *local* rule proposed in [30, 31] to estimate the probability of a conditional in a problem of type 2. It recently ignated a philosophical debate [19, 33, 34]. In this updating context, this rule is axiomatically justified [29, 49].

**Table 1** Focusing and updating contexts for chips and Bertrand's three boxes problem

| Problem | Focusing (Bayes' rule) | Updating (Minimal rule) |
|---|---|---|
| <br>**Type 1.** *A chip is chosen at random. Suppose the chip is square. What are the chances that it is black?* | <br>$P(B\vert S) = \frac{\frac{2}{7}}{\frac{3}{7}} = \frac{2}{3}$<br>(Conditional probability) | <br>1. Remove round chips<br>2. Count $n(B \wedge S)$ on $n(S)$:[a] $P_{\cancel{\neg S}}(B) = \frac{2}{3}$ |
| <br>**Type 2.** *A chip is chosen at random from a box. Suppose the chip is square. What are the chances that box A has been selected?* | <br>$P(B\vert S) = \frac{\frac{1}{3}\times 1}{(\frac{1}{3}\times 1)+(\frac{1}{3}\times\frac{1}{2})} = \frac{2}{3}$<br>(Bayes' identity form) | <br>1. Remove round chips<br>2. Remove the empty box C<br>3. Count $n(A \wedge S)$ on $n(S)$: $P_{\cancel{\neg S}} = \frac{1}{2}$ |

[a] Let $n(x)$ be the number of chips with the $x$ characteristics in the layout.

For the Type 1 problems, both Bayesian and minimal rules give the same solution.[7] However in type 2 problems, whether focusing or updating, the revision contexts imply different computational processes that clearly give two different solutions.

In this paper, our objective is to show that when confronted with type 1 problems, individuals use an updating cognitive strategy and apply the minimal rule. They should also proceed in the same way when evaluating the conditional probability.

[7] The minimal rule is isomorphic to redistributing the weights of removed world(s) proportionally to the remaining world(s). For the type 1 problem, it mathematically corresponds to Bayes' rule:

$$\begin{aligned} P(H|D) &= \frac{P(H \wedge D)}{P(H \wedge D) + P(\neg H \wedge D)} \\ &= \frac{P(H \wedge D)}{P(H \wedge D) + P(\neg H \wedge D)} + P(H \wedge D) \times \left(1 - \frac{P(H \wedge D) + P(\neg H \wedge D)}{P(H \wedge D) + P(\neg H \wedge D)}\right) \\ &= P(H \wedge D) + \frac{P(H \wedge D)}{P(H \wedge D) + P(\neg H \wedge D)} \times P(\neg D) = P_{\cancel{\neg D}}(H) \end{aligned}$$

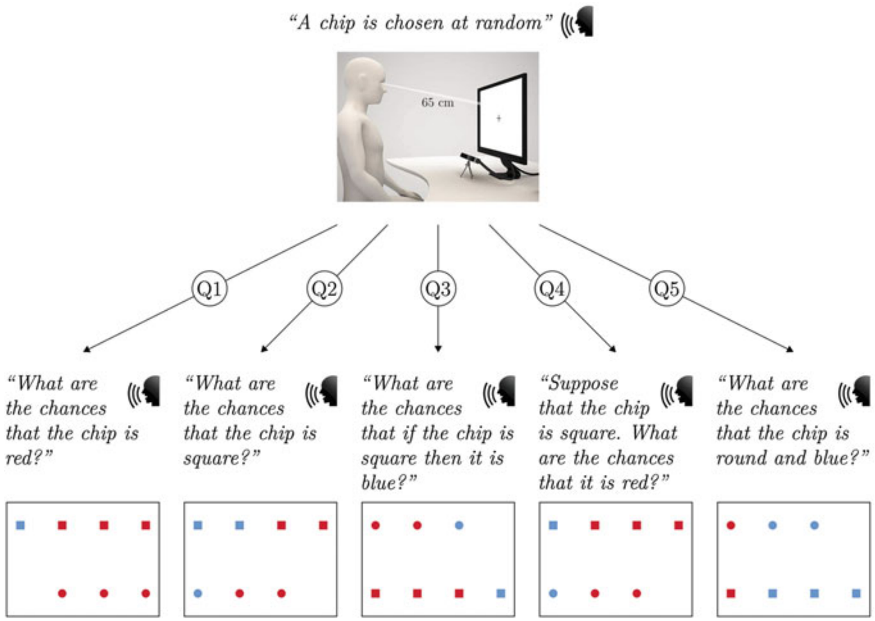

"A chip is chosen at random"
65 cm
Q1
Q2
Q3
Q4
Q5
"What are the chances that the chip is red?"
"What are the chances that the chip is square?"
"What are the chances that if the chip is square then it is blue?"
"Suppose that the chip is square. What are the chances that it is red?"
"What are the chances that the chip is round and blue?"

**Fig. 1** Experiment's procedure for the order $P(\textit{if } S \textit{ then } C)$, $P(C|S)$ and $P(S \wedge C)$

# 4 The Experiment

## *4.1 Methodology*

**Participants**: 19 students (9 female), aged between 20 and 40 (M = 25.33, SD = 3.75). All of them had completed high school and were *native French speakers*. Their background covered all disciplines from 0 to 5 years of higher education with a mean of 3 years. They responded at their own pace and were orally guided through the experiment (see the Fig. 1).

**Apparatus**: The equipment came from the *SMI Eye-tracking Package* (SensoMotoric Instruments GmbH, Teltow, Germany). The experiment was programmed using the *SMI ExperimentCenter 2* v*3.5.144* software. Eye-movement parameters were measured using an *SMI REDm* eye-tracker with a sampling rate of 120 Hz (60 Hz for each eye).

**Procedure**: Seven chips of two colors (blue and red)[8] and of two shapes (round and square) were displayed in two rows as in [47].[9] All participants saw the same five chip

[8] These colors were better discriminated.

[9] The viewing angle of the stimuli was 9.6 on a 1920 × 1080 resolution computer screen.

**Table 2** Average (and median) of participant's time eye-fixations in AOIs (in ms) (N=18)[a]

| AOIs | $P(\textit{if } S \textit{ then } C)$ | $P(C\|S)$ | $P(S \wedge C)$ | TTOCF[b] |
|---|---|---|---|---|
| $S \wedge C$ | 2390 (2366) | 2789 (1894) | 3016 (1829) | 2732 (2030) |
| $S \wedge \neg C$ | 1679 (940) | 1025 (724) | 947 (498) | 1217 (721) |
| $\neg S$ | 48 (0) | 49 (0) | 1829 (1395) | 642 (465) |
| TTOCF[b] | 1332 (1102) | 1288 (873) | 1931 (1241) | |

[a] One participant has been removed because of a bad eye-tracking recording.
[b] Total times of chip fixations.

layouts (one for each of the five questions, Fig. 1). The first two questions ($Q_1$ and $Q_2$) ensured that the participants had understood the instructions. The three following questions ($Q_3$, $Q_4$ and $Q_5$) were randomized[10] and corresponded respectively to evaluations of (with $S$ referring to the shape of the chip and $C$ to the color):

1. the probability of a conditional $P(\textit{if } S \textit{ then } C)$,
2. the conditional probability $P(C|S)$ and
3. the probability of a conjunction $P(S \wedge C)$.

## 4.2 Results and Discussion

90 % of participants answered respectively $\frac{n(S\wedge C)}{n(S\wedge C)+n(S\wedge\neg C)}$ for $P(\textit{if } S \textit{ then } C)$ and for $P(C|S)$. All participants gave $\frac{n(S\wedge C)}{n(S\wedge C)+n(S\wedge\neg C)+n(\neg S)}$ for $P(S \wedge C)$. These results are very close to the results of [47]. The eye-fixation time in *areas of interest* (AOIs) were defined for $S \wedge C$, $S \wedge \neg C$ and $\neg S$ for all groups and the eye-fixation time was calculated[11] (see Table 2).

A Friedman's test shows that the data are not distributed in the same way in the three experimental conditions only for the $\neg S$ AOI (Friedman's $\chi^2 = 32.54$, $df = 2$, $p - value < 0.0001$). The multiple comparison analysis shows that the absolute value of the mean rank sums difference of two conditions $P(\textit{if } S \textit{ then } C)$ and $P(C|S)$ is not significant ($|\bar{R}_{P(\textit{if } S \textit{ then } C)} - \bar{R}_{P(C|S)}| = 1 < 14.36$). However for both conditions the median difference is significant with the $P(S \wedge C)$ condition (respectively $|\bar{R}_{P(\textit{if } S \textit{ then } C)} - \bar{R}_{P(S\wedge C)}| = 26.5 > 14.36$ and $|\bar{R}_{P(C|S)} - \bar{R}_{P(S\wedge C)}| = 27.5 > 14.36$). These results are consistent with our hypothesis. To evaluate $P(\textit{if } S \textit{ then } C)$ and $P(C|S)$ participants have only looked at $S \wedge C$ and $S \wedge \neg C$ AOIs. However for $P(S \wedge C)$, participants have also looked the $\neg S$ AOI.

[10] To avoid a possible order effect, the participants were randomly allocated to six groups which all answered five questions (two controls and the three conditionals in different orders).

[11] The gaze behavior were also recorded for scan path and visual strategy (VS) investigation.

## 5 Conclusion

In this experiment aiming at analysing the Equation, applying an eye-tracking methodology, we find that participants evaluate the probability of a conditional $P(\textit{if } A \textit{ then } C)$ in the same way they evaluate the conditional probability $P(C|A)$. However, for both evaluations, participants tend to naturally consider an updating context of revision. Yet this result rises a concern for the conditioning principle that should be more precisely studied in further experiments.

## References

1. Baratgin J (2009) Updating our beliefs about inconsistency: the Monty-Hall case. Math Soc Sci 57:67–95
2. Baratgin J (2015) Rationality, the Bayesian standpoint, and the Monty-Hall problem. Front Psychol 6:1168
3. Baratgin J (in press) Le raisonnement humain: une approche finettienne [Human reasoning: A Finettian approach]. Hermann, Paris
4. Baratgin J, Douven I, St Evans JBT, Oaksford M, Over D, Polytzer G, (2015) The new paradigm and mental models. Trends Cogn Sci 19:547–548
5. Baratgin J, Over DE, Politzer G (2013) Uncertainty and the de Finetti tables. Think Reason 19:308–328
6. Baratgin J, Over DE, Politzer G (2014) New psychological paradigm for conditionals and general de Finetti tables. Mind Lang 29:73–84
7. Baratgin J, Politzer G (2006) Is the mind Bayesian? The case for agnosticism. Mind Soc 5:1–38
8. Baratgin J, Politzer G (2007) The psychology of dynamic probability judgment: order effect, normative theories and experimental methodology. Mind Soc 6:53–66
9. Baratgin J, Politzer G (2010) Updating: a psychologically basic situation of probability revision. Think Reason 16:253–287
10. Baratgin J, Politzer G (2016) Logic, probability and inference: a methodology for a new paradigm. In: Macchi L, Bagassi M, Viale R (eds) Cognitive unconscious and human rationality. MIT Press, Cambridge
11. Baratgin J, Politzer G, Over DE, Takahashi T (2016) The psychology of uncertainty and three-valued truth tables. MS
12. Coletti G, Scozzafava G (2002) Probabilistic logic in a coherent setting. Kluwer, Dordrecht
13. Cruz N, Oberauer K (2014) Comparing the meanings of "if" and "all". Mem Cogn 42:1345–1356
14. Cruz N, Baratgin J, Oaksford M, Over DE (2015) Bayesian reasoning with ifs and ands and ors. Front Psychol 6:192
15. de Finetti B (1957) L'informazione, il ragionamento, linconscio nei rapporti con la previsione [The information, reasoning, the unconscious in relations with the prediction]. Lindustria 2:3–27
16. de Finetti B (1964) Foresight: Its logical laws, its subjective sources. In: Kyburg H, Smokier, HE (eds) Studies in subjective probability. Wiley, New York (Original work published 1937)
17. de Finetti B (1980) Probabilità [probability]. Enciclopedia X:1146–1187). Einaudi, Torino
18. de Finetti B (1995) The logic of probability. Philos Stud 77:181–190 (Original work published 1936)
19. Douven I (2008) Kaufmann on the probabilities of conditionals. J Philos Logic 37:259–266
20. Douven I, Verbrugge S (2010) The adams family. Cognition 117:302–318

21. Douven I, Verbrugge S (2013) The probabilities of conditionals revisited. Cogn Sci 117:302–318
22. Dubois D, Prade H (1992) Evidence, knowledge, and belief functions. Int J Approx Reason 6:295–319
23. Elqayam S, Over DE (2013) New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. Think Reason 19:249–265
24. Evans JSBT, Handley SJ, Neilens H, Over DE (2007) Thinking about conditionals: a study of individual differences. Mem Cogn 35:1759–1771
25. Evans JSBT, Handley SJ, Over DE (2003) Conditionals and conditional probability. J Exp Psychol Learn Mem Cogn 29:321–355
26. Edgington D (1995) On conditionals. Mind 104:235–329
27. Fugard JB, Pfeifer N, Mayerhofer B, Kleiter GD (2011) How people interpret conditionals: Shifts toward conditional event. J Exp Psychol Learn Mem Cogn 37:635–648
28. Jeffrey R (1991) Matter-of-fact conditionals. PAS (suppl) 65:161–183
29. Katsuno A, Mendelzon A (1992) On the difference between updating a knowledge base and revising it. In: Gärdenfors P (ed) Belief revision. Cambridge University Press, Cambridge
30. Kaufmann S, Winston E, Zutty D (2004a) Local and global interpretations of conditionals. Eighth Symposium on Logic and Language (LOLA 8), Debrecen, Hungary
31. Kaufmann S (2004b) Conditioning against the grain: abduction and indicative conditionals. J Philos Logic 33:583–606
32. Kern-Isberner G (2001) Revising and updating probabilistic beliefs. In: Williams MA, Rott H (eds) Frontiers in Belief Revision. Kluwer Academic, Dordrecht
33. Khoo J (2016) Probabilities of conditionals in context. Linguist Philos 39:1–43
34. Korzukhin T (2016) Probabilities of conditionals in context. a comment on Khoo (2016). Linguist Philos 39:45–49
35. Manktelow KI, Over DE, Elqayam S (2011) Paradigms shift: Jonathan Evans and the science of reason. In: Manktelow KI, Over DE, Elqayam S (eds) The science of reason: A festschrift for J. St. B. T. Evans. Psychology Press, Hove
36. Oberauer K, Wilhelm O (2003) The meaning(s) of conditionals: Conditional probabilities, mental models and personal utilities. J Exp Psychol Learn Mem Cogn 29:680–693
37. Olsen NS, Singmann H, Klauer C (in press) The Relevance Effect and Conditionals. Cogn
38. Over DE (2009) New paradigm psychology of reasoning. Think Reason 15:431–438
39. Over DE, Baratgin J (2016) The 'defective' truth table: Its past, present, and future. In: Lucas E, Galbraith N, Over DE (eds) The thinking mind: the use of thinking in everyday life. Psychology Press, Alberton
40. Over DE, Hadjichristidis C, St Evans JBT, Handley SJ, Sloman SA (2007) The probability of causal conditionals. Cogn Psychol 54:62–97
41. Pfeifer N, Kleiter GD (2006) Inference in conditional probability logic. Kybernetika 42:391–404
42. Pfeifer N, Kleiter GD (2009) Framing human inference by coherence based probability logic. J Appl Log 7:206–217
43. Pfeifer N, Kleiter GD (2010) The conditional in mental probability logic. In: Oaksford M, Chater N (eds) Cognition and conditionals: Probability and logic in human thinking. Oxford University Press, Oxford
44. Pfeifer N, Kleiter GD (2011) Uncertain deductive reasoning. In: Manktelow KI, Over DE, Elqayam S (eds) The science of reason: A festschrift for J. St. B. T. Evans. Psychology Press, Hove
45. Politzer G (in press). Deductive reasoning under uncertainty: A water tank analogy. Erkenntnis
46. Politzer G, Baratgin J (2016) Deductive schemas with uncertain premises using qualitative probability expressions. Think Reason 22:78–98
47. Politzer G, Over DE, Baratgin J (2010) Betting on conditionals. Think Reason 16:172–1777
48. Singmann H, Klauer KC, Over DE (2014) New normative standards of conditional reasoning and the dual-source model. Front Psychol 5:316

49. Walliser B, Zwirn D (2002) Can Bayes rule be justified by cognitive rationality principles? Theory Decis 53:95–135
50. Walliser B, Zwirn D (2011) Change rules for hierarchical beliefs. Int J Approx Reason 52:166–183

# Black-Litterman Model with Multiple Experts' Linguistic Views

**Marcin Bartkowiak and Aleksandra Rutkowska**

**Abstract** This paper presents fuzzy extensions of the Black-Litterman portfolio selection model. Black and Litterman identified two sources of information about expected returns and combined these two sources of information into one expected return formula. The first source of information is the expected returns that follow from the Capital Asset Pricing Model and thus should hold if the market is in equilibrium. The second source of information is comprised of the views held by investors. The presented extension, owing to the use of fuzzy random variables, includes two elements that are important from the point of view of practice: linguistic information and the views of multiple experts. The paper introduces the model extension step-by-step and presents an empirical example.

## 1 Introduction

Modern portfolio theory attempts to maximize a portfolios expected return for a given amount of portfolio risk, or equivalently to minimize the risk for a given level of expected return, by carefully choosing the proportions of various assets. The breakthrough article entitled "Portfolio Selection" was published by Markowitz [19]. The next revolutionary papers in portfolio selection were published by Sharpe [23], Lintner [17] and Black [2]. The model developed by Sharpe and Lintner, known as the Capital Asset Pricing Model (further, CAPM), is still widely used in applications and research studies that deal with risk and returns. The attraction of the CAPM is that it offers powerful and intuitively pleasing predictions about how to measure risk and the relationship between expected return and risk. Unfortunately, the empirical record of the model is poor. Canonical portfolio optimization takes as inputs only the expectations and covariances of a set of assets computed from a given reference

M. Bartkowiak · A. Rutkowska (✉)
Department of Applied Mathematics, Poznan University of Economics and Business,
Poznań, Poland
e-mail: aleksandra.rutkowska@ue.poznan.pl

M. Bartkowiak
e-mail: m.bartkowiak@ue.poznan.pl

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_5

econometric model. The Black-Litterman model (further, BL model), which was first published by Fischer Black and Robert Litterman [3], provides a framework in which more satisfactory results can be obtained from a larger set of inputs: the view portfolios, the expected returns on those portfolios, the confidence in the view portfolios. In other words, the BL model enables investors to combine their unique views regarding the performance of various assets with the market equilibrium by mixing different types of estimates. The BL model was expanded in [4, 5]. The model was discussed in greater detail in [1, 10, 18]. Now, there are a variety of models being labeled as Black-Litterman even though they may be very different from the original model created by Black and Litterman. A comprehensive taxonomy and literature survey was provided in Meucci [20]. Since an investors view of future asset return is always subjective and imprecise, the fuzzy approach seems to be a natural extension of the BL model. Lawrence et al. [16] used fuzzy trapezoidal numbers to represent investor views and omitted the aspect of consistency in combining prior probabilistic distribution and fuzzy views. Gharakhani and Sadjadi [8] assumed views as fuzzy numbers and mean asset return as well as covariance as fixed estimated parameters. They focused on fuzzy compromise programming to find the solution of fuzzy return maximization and fuzzy beta minimization.

In this paper, we introduce extensions of the BL model with linguistic expressed views of future return. As a tool for handling linguistic label information, a fuzzy random variable (further, FRV) is used. FRV concept[1] have been developed in: [14, 15, 22] and in a unified approach in [13]. The overview of different variants can be found in [6, 9]. Operationally, an FRV is a random variable taking fuzzy values. In practice, we are often faced with random experiments whose outcomes are not numbers but are expressed in inexact linguistic terms, in particular, predictions of events in the stock market. Development of the information society resulted in access to a wide range of business and economic information. In many fields of science, there are discussions about taking into account the opinion of many experts at the same time (see [11, 21]). At the moment, the discussion has not moved on in portfolio optimizing literature. Many services such as Bloomberg or Reuters Thompson allow access to the predictions of various experts, and investment firms pay their own experts. Take for example an expert who is questioned about how a planned tax on minerals will affect the valuation of energy companies. Some possible answers would be *a slight decrease in value, no effect, a strong reduction in value* and so on. A natural question is how to take these opinions into account when choosing a portfolio. The BL model allows the inclusion of expert views on the decision process, while an FRV allows the user to calculate: what the average opinion of multiple experts is and how great the uncertainty associated with it is, as well as natural linguistic formulation of expectations.

The paper presents an expert view as the FRV in Puri and Ralescus approach. Since an FRV is a generalization of the random variable, it is possible to combine distributions in the way proposed by Black and Litterman. The main advantage of the

[1] The concept of FRV was introduced by Feron, R., 1976. Ensembles aleatoires flous. C.R. Acad. Sci. Paris, Ser. A (282), pp. 903–906.

BL model is its intuitiveness and ease of application. The extended model presented in this article, despite extensive mathematical tools, keeps these two features.

The rest of this paper is organized as follows: in Sect. 2, we introduce the fundamental theory of FRV and linguistic variables; Sect. 3 presents the new BL model with linguistic views; the next section presents its application to portfolio selection on the Warsaw Stock Exchange; and the last section concludes and summarizes the study.

## 2 Preliminaries

This section briefly reviews some basic concepts of FRV by Puri and Ralescu [22] and linguistic value.

Let $F_0(R^n)$ denote the set of fuzzy subsets: $\mu : R^n \to [0, 1]$ with the following properties:

- $\{x \in R^n : \mu(x) \geq \alpha\}$ is compact for each $\alpha > 0$
- $\{x \in R^n : \mu(x) = 1\} \neq \emptyset$.

For $\mu \in F_0$, $[\mu]^\alpha = \{x \in R : \mu(x) \geq \alpha\}, 0 \leq \alpha \leq 1$ is the $\alpha$-level set of $\mu$ and $\mu^+(\alpha), \mu^-(\alpha)$ are the upper and lower endpoints of $[\mu]^\alpha$.

The addition and scalar multiplication are defined by the following:

$$[\mu + v]^\alpha = [\mu]^\alpha + [v]^\alpha , \tag{1}$$

$$[\lambda\mu]^\alpha = \lambda [\mu]^\alpha , \mu, v \in F_0 \text{ and } \lambda \in R. \tag{2}$$

Operation $\langle ., . \rangle$ is defined by the following equation:

$$\langle \mu, v \rangle = \int_0^1 \left(\mu^-(\alpha) v^-(\alpha) + \mu^+(\alpha) v^+(\alpha)\right) d\alpha. \tag{3}$$

Let $(\Omega, A, P)$ be a probability space where $P$ is a probability measure assumed to be non atomic. A FRV is a function $X : \Omega \to F_0(R^n)$ such that: $\{(\omega, x) : x \in X_\alpha(\omega)\} \in A \times B$, for every $\alpha \in [0, 1]$, where $B$ denotes the Borel subsets of $R^n$, $X_\alpha : \Omega \to P(R^n)$ is define by: $X_\alpha(\omega) = \{x \in R^n : X(\omega)(x) \geq \alpha\}$.

The expected value of a FRV defined by Puri and Ralescu [22] is the Aumann-type mean, which extends the mean of a real-valued d preserving its main properties and behavior. The expected value is a fuzzy number, but the variance and covariance of FRVs[2] according to Feng et al. [7] is scalar, which determines the spread or dispersion of the FRV around its expected value.

The expected value of $X$, denoted by $E(X)$ is the fuzzy set $\mu \in F_0(R^n)$ such that $\{x \in R^n : \mu(x) \geq \alpha\} = \int X_\alpha$ for every $\alpha \in [0, 1]$, where $\int X_\alpha = \left\{\int_\Omega f dP : f \in S\right.$

[2] Variance of FRV have several definition of variance (cf. [6]).

$(X_\alpha)\}$ is the Aumann integral of $X_\alpha$ with respect to $P$ and $S(F)$ is a nonempty bounded set with respect to the $L^1(P)$-norm.

If $\mu_i : R \to [0,1], i = 1, 2, \ldots, n$ are continuous with compact support and $P(X = \mu_i) = p_i$, then

$$E(X) = \Sigma_{i=1}^{n} p_i \mu_i \tag{4}$$

Feng [7] proposed the following equation to calculate covariance and variance of FRV $X, Y$:

$$Cov(X,Y) = \frac{1}{2}\left(E\langle X, Y\rangle - \langle EX, EY\rangle\right), \tag{5}$$

$$DX = Cov(X, X). \tag{6}$$

Thus:

$$DX = \frac{1}{2}\left(E\langle X, X\rangle - \langle EX, EX\rangle\right) = \frac{1}{2}\left(\Sigma_{i=1}^{n} p_i \langle \mu_i, \mu_i\rangle - \Sigma_{i=1}^{n}\Sigma_{j=1}^{n} p_i p_j \langle \mu_i, \mu_j\rangle\right). \tag{7}$$

Further, in the paper we will use FRV with triangular membership function. The triangular fuzzy number is a special type of $LR$ fuzzy number—represented by its core $a$ (most likely value), where left $\alpha$ and right $\beta$ spread (lower and upper bounds) and the notation $(a, \alpha, \beta)$ is used for linear shape functions.

The concept of a linguistic variable was introduced by Zadeh [24]. A linguistic variable is characterized by a quintuple $(\sigma, T(\sigma), U, G, A)$ where $\sigma$ is the name of the variable, $T(\sigma)$ is the set of terms of $\sigma$, $U$ is the universe of discourse, $G$ is a syntactic rule for generating the labels in the terms set, and $A$ is the semantic $U$ rule for associating the meaning to each element of $T(\sigma)$.

## 3 Black-Litterman Model with Linguistic Views

This section introduces the BL model with linguistic views. The BL model starts with a neutral equilibrium portfolio for the prior estimate of returns, because the model relies on General Equilibrium Theory. This part of the BL model is in the new approach without any changes. Investors have specific views regarding the expected return of some of the assets in a portfolio, which differ from the implied equilibrium return. The BL model allows such views to be expressed in either absolute or relative terms. The new BL model also allows them to be expressed in linguistic form and allows the occurrence of many predictions/views in relation to a single asset to be shown. For example, the standard BL format view would look like: *Banks will have an absolute excess return of -5 %.*After fuzzy modification in the BL model, we can consider multiple views in the following form: expert 1: *Banks will lose slightly*, expert 2: *The situation will not affect the bank's valuation*, expert 3: *The Eurozone banking system may lose value about 3 %.*

We assume that there is $m$ experts and that every investor view is an FRV with discrete probability and continuous membership functions, which describe the linguistic term set. So, every view is represented by two vectors: $\tilde{\mu}$ the vector of fuzzy sets describes corresponding linguistic terms; $p$ the vector of view confidence. In the proposed solution, vector $p$ is responsible for determining certain weight in relation to the opinions of many experts.[3] For every view described by $\mu$ and $p$ according to formula (4), we calculate the expected value. During the decision-making process, the investor has a collection of views. We will represent the investors $k$ views on $n$ assets in analogy to the BL model, using the following matrices:

- $P$, a $k \times n$ matrix of the asset weights within each view. The matrix is the same as in the standard BL model.
- $\tilde{Q}$, a $k \times 1$ fuzzy vector of the fuzzy expected returns for each view.
- $\Omega$, a $k \times k$ matrix of the covariance according to formula (7) of the views. $\Omega$ is diagonal as it will require each view to be unique and uncorrelated with the other views.

After the specification of the prior estimate of returns $(\pi, \Sigma)$, the scalar $\tau$, the fuzzy views $\tilde{Q}$ and the covariance matrix of the error $\Omega$ all of the inputs are then entered into the BL formula, and the new combined return vector $E\tilde{[r]}$ is derived as follows:

$$\tilde{E}[r] = \left[(\tau\Sigma)^{-1} + P^{'}\Omega^{-1}P\right]^{-1}\left[(\tau\Sigma)^{-1}\pi + P^{'}\Omega^{-1}\tilde{Q}\right] \tag{8}$$

The covariance matrix of the joint distribution is:

$$\tilde{M} = \left[(\tau\Sigma)^{-1} + P^{'}\Omega^{-1}P\right]^{-1} \tag{9}$$

Optimal portfolio weights are computed by solving the optimization problem. It can be a traditional mean-variance approach starting from equilibrium expected returns as well as the maximization of the utility function. The same as when computing the equilibrium returns (4), we will use the following quadratic utility function:

$$\tilde{U} = w^T\tilde{E} - \frac{\delta}{2}w^T\tilde{M}w \tag{10}$$

where $w$ is the vector of weights invested in each asset, $\tilde{E}$—the new combined return vector, $\tilde{M}$—new covariance matrix. As expected returns are fuzzy vectors, it is a fuzzy optimization problem. A different method for fuzzy optimization can be found in [12]. Arriving at the optimal portfolio is somewhat more complex in the presence of constraints. We only consider budget constraint which forces the sum of the total portfolio weights to be one.

[3] The aggregation operator can be considered as a separate research issue. This paper illustrates a new BL algorithm; so, vector $p$ is set in the simplest way through the frequency of the experts answers, with the assumption that all the opinions are equivalent.

**Table 1** Fuzzy relation of the linguistic variable *Informations influence on a share*

| $t_i$ | $\mu_i$ |
|---|---|
| A significant increase | (0.1, 0.04, 0.05) |
| An increase | (0.05, 0.03, 0.03) |
| A slight increase | (0.01, 0.01, 0.015) |
| No influence | (0, 0.005, 0.005) |
| A slight decrease | (−0.01, 0.015, 0.015) |
| A decrease | (−0.05, 0.03, 0.03) |
| A significant decrease | (−0.1, 0.05, 0.04) |

## 4 Empirical Study

The goal of this section is to test the new BL model and compare the results with those of the standard BL model. To implement our study, we selected all the stocks from the WIG20 index.[4] The data series starts from 2010 to 2014, with daily observations. This time period includes the 2011 European debt crisis, 2012–2013 stock boom and stagnation. The set of terms of *information influence* with the assignment of a meaning to each label is presented in Table 1. At the beginning of each half-year, we calculate the vector of equilibrium excess return and the covariance matrix of the excess returns. For these calculations, we use data from the previous year. We calculate the risk aversion parameter for Polish equity markets to be 3.78 and we use $\delta = 4$ for our model; although, our results are not materially affected by this choice of parameter. Then we find the maximum of function with the following constraints: $(i)$ the sum of weights is equal 1, $(ii)$ there is no short selling. In this way, we find two equilibrium portfolios that are used as a benchmark. Next, we define views (3 views on each half-year), both in a non-fuzzy and linguistic way. Afterwards, we again optimize the utility function and fuzzy utility function (10) with constraints $(i)$ and $(i) + (ii)$. In both cases, we use $\tau = 1$ and impose an additional condition that the weight of a single share does not exceed 20 %. This additional constraint ensures adequate portfolio diversification. As mentioned above, the described algorithm was repeated every six months, i.e. we rebuild a portfolio with semiannual frequency. To implement the same algorithm in both cases, in the fuzzy case centroid defuzzification of the objective function was first done and then optimization. Table 2 shows the performance of portfolios from January 2011 to December 2014. All BL portfolios outperformed the market portfolio (i.e. the WIG20 index) and equilibrium portfolios. Returns from the fuzzy BL portfolios exceed the returns from the canonical BL portfolios. In the case of the portfolio with constraint $(i)$, the difference is not large: it is less than 0.5 p.p. for the annualized return. However, for the portfolio with all

[4] The WIG20 index is based on the value of a portfolio with shares in the 20 major and most liquid companies on the Warsaw Stock Exchange Main List.

**Table 2** Results: performance of different portfolios from January 2011 to December 2014

| Date | Market | Eq. (i) | Eq. (i) + (ii) | BL (i) | BL (i) + (ii) | fBL (i) | fBL (i) + (ii) |
|---|---|---|---|---|---|---|---|
| 03.01.11 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| 30.06.11 | 10070.77 | 12171.41 | 10863.31 | 10497.29 | 11166.77 | 11369.53 | 11235.6 |
| 30.12.11 | 7707.52 | 8925.14 | 7882.61 | 11781.23 | 10231.61 | 12678.85 | 10294.68 |
| 29.06.12 | 8177.71 | 7418.34 | 7830.23 | 15798.01 | 10668.32 | 16983.07 | 10935.5 |
| 28.12.12 | 9283.55 | 7575.94 | 7563.89 | 20256.2 | 13097.96 | 23623.72 | 13425.99 |
| 28.06.13 | 8071.11 | 6628.58 | 6585.02 | 16470.45 | 12082.06 | 23885.26 | 12391.18 |
| 30.12.13 | 8629.42 | 8490.77 | 8133.46 | 16734.85 | 12501.56 | 27316.71 | 15045.7 |
| 30.06.14 | 8657.56 | 8221.45 | 7887.49 | 22696.59 | 13628.68 | 29591.36 | 15045.7 |
| 30.12.14 | 8323.77 | 7952.04 | 7658.35 | 27234.18 | 13049.75 | 27467.93 | 15334.65 |
| overall return | -16.76% | -20.48% | -23.42% | 172.34% | 30.50% | 174.68% | 53.35% |
| annualized return | -5.93% | -7.35% | -8.51% | 39.65% | 9.28% | 40.05% | 15.32% |

the constraints, the fuzzy BL model generates a portfolio that is almost twice as good as the common BL model.

## 5 Conclusion

The paper presets the fuzzy BL model extension with linguistic views for many view sources. To model the linguistic view, FRV is used. This approach allows intuitive formulating views, as well as the setting of opinions from a group of experts. The empirical tests suggest that fuzzy extensions of the BL model have investment value.

**Acknowledgments** This work was supported by the Polish National Science Center under Grant 2013/09/N/HS4/03761.

## References

1. Bevan A, Winkelmann K (1998) Using the Black-Litterman global asset allocation model: three years of practical experience, Goldman, Sachs & Company
2. Black F (1972) Capital market equilibrium with restricted borrowing. J Bus 45(3):444–455
3. Black F, Litterman R (1990) Asset allocation: combining investors views with market equilibrium. Fixed Income Research, Goldman, Sachs & Company
4. Black F, Litterman R (1991) Global asset allocation with equities, bonds and currencies, Goldman, Sachs & Company
5. Black F, Litterman R (1992) Global portfolio optimization. Financ Anal J, 28–43
6. Couso I, Dubois D (2009) On the variability of the concept of variance for fuzzy random variables. IEEE Trans Fuzzy Syst 5(17):1070–1080
7. Feng Y, Hu L, Shu H (2001) The variance and covariance of fuzzy random variables and their applications. Fuzzy Sets Syst 120:487–497
8. Gharakhani M, Sadjadi S (2013) A fuzzy compromise programming approach for the Black-Litterman portfolio selection model. Decis Sci Lett 1(2):11–22
9. Gil M, Lopex-Diz M, Ralescu D (2006) Overview on the development of fuzzy random variables. Fuzzy Sets Syst 157:2546–2557
10. He G, Litterman R (1999) The intuition behind Black-Litterman model portfolios, Goldman, Sachs & Company
11. Huynh V-N, Nakamori Y (2005) Multi-expert decision-making with linguistic information: a probabilistic-based model. IEEE, Hawaii
12. Kacprzyk J, Orlovski S (2013) Optimization models using fuzzy sets and possibility theory. Springer, Netherlands
13. Krätschmer V (2001) A unified approach to fuzzy variables. Fuzzy Sets Syst 123:1–9
14. Kruse R, Meyer KD (1987) Statistic with vague data. Reidel Publishing Company
15. Kwakernaak H (1989) Fuzzy random variables. definition and theorems. Inf Sci 15:1–29
16. Lawrence K, Pai DR, Klimberg RK, Lawrence SM (2009) A fuzzy programming approach to financial portfolio model. Financial Modeling Applications and Data Envelopment Applications 53
17. Lintner J (1965) The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Rev Econ Stat 47(1):12–37
18. Litterman R, the Quantitative Resources Group (2003) Modern investment management: an equilibrium approach. John Wiley & Sons, New Jersey

19. Markowitz HM (1952) Portfolio selection. J Financ 7(1):77–91
20. Meucci A (2008) The Black-Litterman approach: original model and extensions. http://ssrn.com/abstract=1117574
21. Noor-E-Alama M, Ferdousi Lipi T, Ahsan Akhtar Hasin M, Ullah A (2011) Algorithms for fuzzy multi expert multi criteria decision making (ME-MCDM). Knowledge-Based Systems, April, 367–377
22. Puri ML (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
23. Sharpe W (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19(3):425–442
24. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning—I. Inf Sci 8:199–249

# Representing Lightweight Ontologies in a Product-Based Possibility Theory Framework

**Salem Benferhat, Khaoula Boutouhami, Faiza Khellaf and Farid Nouioua**

**Abstract** This paper investigates an extension of lightweight ontologies, encoded here in DL-Lite languages, to the product-based possibility theory framework. We first introduce the language (and its associated semantics) used for representing uncertainty in lightweight ontologies. We show that, contrarily to a min-based possibilistic DL-Lite, query answering in a product-based possibility theory is a hard task. We provide equivalent transformations between the problem of computing an inconsistency degree (the key notion in reasoning from a possibilistic DL-Lite knowledge base) and the weighted maximum 2-Horn SAT problem.

## 1 Introduction

Knowledge representation for the semantic web requires an analysis of the universe of discourse in terms of concepts, definitions, objects, roles, etc., and then selecting a computer-usable version of the results. Ontologies play an important role for the success of the semantic web as they provide shared vocabularies for different domains, such as medicine and bio-informatics. There are many representation languages for ontologies. Among them, description logics [2] provide solid theoretical foundations to ontologies thanks to their clear semantics and formal properties. Moreover, despite its syntactical restrictions, the DL-Lite family enjoys good computational properties while still offering interesting capabilities in representing terminological knowledge

S. Benferhat · K. Boutouhami
CRIL - CNRS UMR 8188, University of Artois, 62307 Lens, France
e-mail: benferhat@cril.univ-artois.fr

K. Boutouhami · F. Khellaf
RIIMA, University of Sciences and Technology Houari Boumediene,
Bab Ezzouar, Algeria

F. Nouioua (✉)
Aix-Marseille University, CNRS, ENSAM, University of Toulon, LSIS UMR 7296,
Marseille, France
e-mail: farid.nouioua@lsis.org

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_6

[1]. This is why a large amount of works has been recently dedicated to this family and this paper is a contribution to this general research line.

The dynamic of information available on the web naturally leads to a continuous evolution of ontologies and to a permanent need to merge or to align them. As a result, we are often confronted to uncertainties in the used information. Proposing efficient methods for handling uncertainty in description logics, and particularly in the DL-Lite family, is an important research topic. Several recent works are devoted to fuzzy extensions of description logics (see e.g. [4, 8, 10]). Other works propose (min-based) possibilistic extensions of description logics and focus on standard reasoning services (see e.g. [3, 6, 7, 9]).

In some applications the nature of the encountered uncertainty is quantitative. This paper investigates the product-based possibilitic DL-Lite, denoted by Pb-$\pi$-DL-Lite, which has not been considered before. This paper shows that contrarily to the min-based possibilistic DL-Lite, query answering from a Pb-$\pi$-DL-Lite knowledge base is no longer tractable. We provide an encoding of computing the inconsistency degree of the product-based possibilistic DL-Lite knowledge base (the basis of query answering in Pb-$\pi$-DL-Lite) using a weighted maximum 2-Horn satisfiability problem.

## 2 Product-Based Possibilistic DL-Lite

DL-Lite is a family of DLs that aims to capture some of the most popular conceptual modeling formalisms. A DL KB $K = \langle T, A\rangle$ consists of a set $T$ of concept and role axioms (TBox) and a set $A$ of assertional facts (ABox). In this paper, we only consider the DL-Lite$_{core}$ and DL-Lite$_R$ that underlie OWL2-QL [5]. The syntax of the DL-Lite$_{core}$ language is defined as follows:

$$\begin{array}{ll} B \rightarrow A|\exists R & C \rightarrow B|\neg B \\ R \rightarrow P|P^- & E \rightarrow R|\neg R \end{array} \tag{1}$$

where $A$ denotes an atomic concept, $P$ an atomic role, $P^-$ the inverse of the atomic role $P$, $B$ (resp. $C$) are called basic (resp. complex) concepts and roles $R$ (resp. $E$) are called basic (resp. complex) roles.

A DL-Lite$_{core}$ TBox is a set of inclusion axioms of the form: $B \sqsubseteq C$. An ABox is a set of membership assertions on atomic concepts and on atomic roles of the form: $A(a)$ and $P(a, b)$ respectively, where $a$ and $b$ are two individuals.

The DL-Lite$_R$ language extends DL-Lite$_{core}$ with the ability of specifying in the TBox inclusion axioms between roles of the form:

$$E \sqsubseteq R$$

## 2.1 Weighted Assertional DL-Lite Knowledge Base

The syntax of product-based possibilistic DL-lite is represented by the concept of a Pb-$\pi$-DL-Lite knowledge base KB denoted by $K$.

**Definition 1** A Pb-$\pi$-DL-Lite KB $K = \{\langle \phi_i, \alpha_i \rangle : i = 1, \ldots, n\}$ is a finite set of possibilistic axioms of the form $\langle \phi_i, \alpha_i \rangle$, where $\phi_i$ is an axiom expressed in DL-Lite and $\alpha_i \in ]0, 1]$ represents the certainty degree of $\phi_i$.

In Definition 1, only somewhat certain facts (having certainty degrees $> 0$) are considered. We consider that all the TBox axioms are fully certain. This means that there is no uncertainty about the general relationships between concepts and roles, but only about the ABox assertions. Hence, the terminological base is assumed to be stable and should not be questioned in the presence of inconsistencies.

*Example 1* Let us consider the Pb-$\pi$-DL-Lite KB $K$ composed of the following TBox $T$ and ABox $A$, which will be used in the rest of the paper.

$$\begin{aligned} T = \{ \; & \langle Supervisor \sqsubseteq \neg PhD\_Stud, 1.0 \rangle, \langle \exists Supervision\_a \sqsubseteq Supervisor, 1.0 \rangle, \\ & \langle \exists Supervision\_a^{-} \sqsubseteq PhD\_Stud, 1.0 \rangle \}. \\ A = \{ \; & \langle Supervisor(b), 0.11 \rangle, \langle Supervisor(h), 0.04 \rangle, \langle PhD\_Stud(a), 0.19 \rangle, \\ & \langle Supervision(a, b), 0.89 \rangle, \langle Supervision(a, h), 0.30 \rangle \}. \end{aligned}$$

## 2.2 Semantics

A possibility distribution is a function that assigns to each DL-lite interpretation $I$, a real number in the interval [0, 1], called a possibility degree. $\pi_K(I)$ represents the degree of compatibility of $I$ with respect to the available information given in K. If an interpretation $I$ is a model of each axiom of $T$ and each assertion of $A$ then its possibility degree is equal to 1. This reflects the fact that $I$ is fully compatible with $\langle T, A \rangle$. It also obviously means that $\langle T, A \rangle$ is consistent. Now, if $I$ falsifies some axioms of $T$ or some fully certain assertions of $A$, then its possibility degree is equal to 0. This reflects the fact that $I$ is impossible and should not be considered in the query answering process. More generally, if an interpretation $I$ falsifies some assertions of the ABox, then its possibility degree is inversely proportional to the product of the weights of the assertions that it falsifies.

**Definition 2** For all $I \in \Omega$,

$$\pi_K(I) = \begin{cases} 1 & if \;\; \forall \langle \phi_i, \alpha_i \rangle \in K, I \models \phi_i \\ *\{1 - \alpha_i : \langle \phi_i, \alpha_i \rangle \in K, I \not\models \phi_i\} & \text{otherwise} \end{cases} \tag{2}$$

where $\models$ is the satisfaction relation between DL-lite interpretations and DL-Lite formulas.

*Example 2* Let us consider again Example 1. The following table gives an example of the possibility degrees, obtained using Definition 2, for four interpretations over the domain $\triangle = \{a, b\}$.

| $I$ | $.^I$ | $\pi_K(I)$ |
|---|---|---|
| $I_1$ | $PHD^I = \{a\}$, $Super^I = \{b, h\}$, $Supervision\_a^I = \{(a, b), (a, h)\}$ | 0.000 |
| $I_2$ | $PHD^I = \{a\}$, $Super^I = \{b, h\}$, $Supervision\_a^I = \{(b, a), (h, a)\}$ | 0.077 |
| $I_3$ | $PHD^I = \{a\}$, $Super^I = \{b, h\}$, $Supervision\_a^I = \{(a, b), (a, h)\}$ | 0.000 |
| $I_4$ | $PHD^I = \{b, h\}$, $Super^I = \{a\}$, $Supervision\_a^I = \{(a, b), (a, h)\}$ | 0.692 |

A Pb-$\pi$-DL-Lite KB $K$ is said to be fully consistent if there exists an interpretation $I$ such that $\pi_K(I) = 1$. Otherwise, $K$ is said to be somewhat inconsistent. In the presence of certainty degrees associated with assertions, the concept of inconsistency becomes a graduated notion. More formally:

**Definition 3** Let $K$ be a Pb-$\pi$-DL-Lite KB and $\pi_K$ be the possibility distribution induced by $K$ obtained by Eq. 2. The inconsistency degree of $K$, denoted by $Inc(K)$, is semantically defined as follows:

$$Inc(K) = 1 - \max_{I \in \Omega}(\pi_K(I)) \tag{3}$$

*Example 3* The inconsistency degree of the KB $K$, presented in Example 1 is: $Inc(K) = 1 - \max_{I \in \Omega}(\pi_K(I)) = 0.30$.

# 3 Inconsistency Degree as a Weighted Max-2-Horn-SAT Problem

## 3.1 Weighted Max-2-Horn-SAT (WM2HSAT)

The WM2HSAT problem consists in finding an assignment of boolean values to the propositional variables that maximizes the total weights of satisfied clauses. A weighted 2-Horn KB is a set of weighted formulas of the form : $\theta = \{(\phi_i, k_i) : i = 1, ..., n\}$ where $\phi_i$ is a clause with at most two literals and at most one non-negative literal. $k_i$ is a natural number belonging to $N \cup \{\infty\}$.

**Definition 4** Let $\theta$ be a weighted 2-Horn KB and $R$ be a positive integer. The weighted Max-2-Horn-SAT decision problem is defined as follows:

> Is there a sub-base $\theta' \subseteq \theta$ such that i) $\{\phi_i : (\phi_i, k_i) \in \theta'\}$ is consistent, ii) $\theta'$ contains every $(\phi, k)$ of $\theta$ such that $k = \infty$ and iii) $Weight(\theta') \geqslant R$?

where $Weight(\theta') = \sum\{k_i : (\phi_i, k_i) \in \theta' \text{ and } k_i \neq \infty\}$.

Note that maximizing $Weight(\theta')$ is equivalent to minimizing $Cost(\theta')$ given by: $Cost(\theta') = \sum\{k_i : (\phi_i, k_i) \in \theta \setminus \theta' \textit{ and } k_i \neq \infty\}$.

## 3.2 *From an Inconsistent Pb-π-DL-Lite KB to a Weighted 2-Horn KB*

A conflict is a minimal sub-base of $A$ which is inconsistent with $A$. A set of assertions $\mathcal{C}$ is said to be a conflict if: i) $\mathcal{C} \subseteq A$, ii) $\langle T, C\rangle$ is inconsistent and iii)$\forall \mathcal{C}' \subseteq \mathcal{C}, T \cup \mathcal{C}'$ is consistent. It has been shown in [1] that conflicts in a DL-Lite KB are composed of at most two ABox assertions.

The first step in computing $Inc(K)$ is to encode the set of all conflicts and their corresponding weights by a weighted 2-Horn KB $B_K$.

**Definition 5** Let $K = \langle T, A\rangle$ be a Pb-π-DL-Lite KB. Let $\zeta$ be the set of all conflicts in A. Let $M$ be a sufficiently large integer number. Let $F$ be a scale changing function defined by: $F(x) = -10^M * (ln(1-x))$. Each assertional fact $X(a)$ is associated with a propositional symbol simply denoted by $X_a$.

The weighted propositional 2-Horn KB corresponding to $K$, denoted by $\mathbb{B}_K$, is defined as follows:
$\mathbb{B}_K = \{(D_a, F(\alpha)), (B_a, F(\beta)), (\neg D_a \vee \neg B_a, \infty) \mid \{(D(a), \alpha), (B(a), \beta)\} \in \zeta\}$.

The function $F$ is not unique. Recall that weights in $K$ are expressed using the unit interval [0, 1] while the weights in $\mathbb{B}_K$ are integers.

*Example 4* The weighted 2-Horn KB corresponding to the Example 1 is:

$\mathbb{B}_K = \{(Supervisor\_b, 11653), (Supervisor\_h, 4082)$ ,
$(PhD\_Stud\_a, 21072), (Superv\_a\_b, 220727), (Superv\_a\_b, 35667),$
$(\neg Superv\_a\_h \vee \neg PhD\_Stud\_a, \infty), (\neg Superv\_a\_h \vee \neg Supervisor\_h, \infty),$
$(\neg Superv\_a\_b \vee \neg PhD\_Stud\_a, \infty), (\neg Superv\_a\_b \vee \neg Supervisor\_b, \infty)\}.$

**Proposition 1** *Let $K$ be a Pb-π-DL-Lite KB and $\mathbb{B}_K$ be its associated weighted propositional 2-Horn KB. $Inc(K) = \alpha$ if and only if there exists a consistent sub-base $\mathbb{B}'_K \subseteq \mathbb{B}_K$ such that i) $Cost(\mathbb{B}'_K) = F(\alpha)$ and ii) for every consistent sub-base $\mathbb{B}''_K \subseteq \mathbb{B}_K$, $Cost(\mathbb{B}''_K) \geq Cost(\mathbb{B}'_K)$.*

Proposition 1 is important since it shows that the inconsistency degree of a Pb-π-DL-Lite KB can be redefined using the cost of a solution of the WM2HSAT problem on the associated weighted propositional KB. This gives us a practical mean to compute inconsistency degree using a WM2HSAT solver.

For the sake of simplicity, we assume that full inconsistency cannot occur, namely we assume that $Inc(K) < 1$.

**Assumption 1** Let $K = \langle T, A\rangle$ be a Pb-π-DL-Lite KB. Then, we assume that $T \cup \{(f, 1) : (f, 1) \in A\}$ is consistent.

**Algorithm 1** Inconsistency degree (T, A)

```
Require: $K = \langle T, A\rangle$: a Pb-$\pi$-DL-Lite knowledge base.
Ensure: $Inc\_K$ {The inconsistency degree of $K$}
  $\zeta = Compute\_Conflicts(T, A)$
  $\mathbb{B}_K = Tansformation(\zeta, A)$
  $l \leftarrow Min\{k_i : (\varphi_i, k_i) \in \mathbb{B}_K\}$
  $u \leftarrow \sum\{k_i : (\varphi_i, k_i) \in \mathbb{B}_K \; and \; k_i \neq \infty\}$
  while ($l < u$) do
    $r \leftarrow (l + u)/2$
    if ($Weighted\_Max\_Horn\_2\_Sat(\mathbb{B}_K, r) =$ True) then
      $u \leftarrow r - 1$
    else
      $l \leftarrow r + 1$
    end if
  end while
  return $(1 - e^{-r \div 10^M})$;
```

The following algorithm accepts as input a Pb-$\pi$-DL-Lite KB and returns its inconsistency degree.

*Example 5* Let us consider the KB $K$ from Example 1. Its corresponding weighted 2-Horn KB $\mathbb{B}_K$ is given in Example 4. The next step consists in a dichotomic search in the interval ranging from the minimum value $l = 4082$ and the maximum value $u = 293201$. The solver WM2HSAT is invoked with the sub-base $\mathbb{B}_K$. The last call of the solver returns the consistent sub-base $\mathbb{B}'_K$ which minimizes the sum of the degrees of formulas outside $\mathbb{B}'_K$ $cost(\mathbb{B}'_K) = 36807$. The last step consists in computing $Inc(K) = 1 - e^{(-36807 \div 10^5)} = 0.30$.

## 4 Query Answering in a Product-Based Possibilistic DL-Lite

The problem of standard query answering is closely related to the ontology-based data access problem which takes as inputs a set of assertions, an ontology and a conjunctive query $q$ and aims to find all answers to $q$ over the set of data. We will limit ourselves to boolean queries. This is not a restriction since a conjunctive query can be equivalently redefined from a family of boolean queries, each of them is a result of instantiating the vector of distinguished variables.

A basic boolean query is called a grounded query, has the form:
$q \leftarrow \exists \overrightarrow{y} \bigwedge_{i=1}^{n} B_i(\overrightarrow{y_i})$, where $B_i$ is either an atomic concept or an atomic role or an individual and $\overrightarrow{y_i}$ is either a variable (if $B_i$ is a concept) or a pair of variables or a variable and an individual (if $B_i$ is a role). Given a boolean query q, we first need to define the concept of a necessity measure, defined by:

$$N(q) = 1 - \max\{\pi_K(I) : I \not\models q\}. \tag{4}$$

$N(q)$ represents to what extent $q$ is certain given the available knowledge.

If $\pi_K(I)$ is fully consistent, then $N(q) > 0$ holds (namely $q$ is somewhat accepted) if and only if each model of axioms of $T$ and assertions of $A$ is also a model of q. Similarly, $N(q) = 1$ (q is fully accepted) if and only if for all $I$ such that $I \not\models q$ we have $\pi(I) = 0$ (namely, all counter models of $q$ are declared as impossible). Now, when $\pi$ is sub-normalized or inconsistent, then $q$ is said to be somewhat accepted if and only if $N(q) > Inc(K)$.

**Definition 6** Let $K$ be a Pb-$\pi$-DL-Lite KB, $\pi_K$ be the possibility distribution associated with $K$ using Eq. 2. Let $N_\pi$ be the necessity measure induced by $\pi_K$ using Eq. 5. Let $q$ be a boolean query. Then $K \models_\pi q$ if and only if $N_\pi(q) > Inc(K)$.

Query answering process comes down first to the reformulation of the query $q$ over the TBox in order to enrich it while eliminating all redundancies using the algorithm $PerfectRef$ proposed in [5]. This step leads to obtain a set of queries $Q$ where the union of the answer sets of these queries will be the answer of the initial query. Hence, querying $q$ comes down to evaluate each query $q_i \in Q$.

A very basic case in query answering is instance checking. The instance checking problem, in standard DL-Lite consists in deciding, given an individual $a$ (or a pair of individuals $(a, b)$) a concept $B$ or a role $R$ and a DL-Lite KB $K = \langle T, A\rangle$, whether $B(a)$(resp. $R(a, b)$) follows from $\langle T, A\rangle$.

**Proposition 2** *Let $K = \langle T, A\rangle$ be a Pb-$\pi$-DL-Lite KB, $B$ be a concept (resp. $R$ be a role) and $a, b$ be two individuals. $D_B$ (resp. $D_R$) is an atomic concept (resp. an atomic role) not appearing in $T$. Then :*

1 *$N(B(a)) = Inc(K_1)$ (resp. $N(R(a, b)) = Inc(K_1)$) where $K_1 = \langle T_1, A_1\rangle$ with $T_1 = T \cup \{(D_B \sqsubseteq \neg B, 1)\}$ (resp. $T_1 = T \cup \{(D_R \sqsubseteq \neg R, 1)\}$) and $A_1 = A \cup \{(D_B(a), 1)\}$ (resp. $A_1 = A \cup \{(D_R(a, b), 1)\}$).*

2 *$B(a)$ (resp. $R(a, b)$) is a consequence of $K$, denoted by $K \models_\pi B(a)$ (resp. $K \models_\pi R(a, b)$) if $Inc(K_1) > Inc(K)$.*

Proposition 2 shows how to evaluate the query by using the concept of inconsistency degree. If a query is composed of a conjunction of assertions (grounded queries), then its enough to apply Proposition 2 to each assertion. This is possible thanks to the following propriety of necessity measures:

$$N(q_1 \wedge q_2) = \min(N(q_1), N(q_2)). \tag{5}$$

## 5 Conclusions and Future Work

This paper developed an extension of lightweight ontologies, encoded in DL-Lite language, to the product-based possibility theory framework. The resulting language is denoted Pb-$\pi$-DL-Lite. The paper first introduced the syntax and the semantics of Pb-$\pi$-DL-Lite. Then, it addressed the problem of query answering and showed that it comes down to the problem of computing inconsistency degree in product-based

Pb-$\pi$-DL-Lite knowledge bases. This problem is intractable in product-based DL-Lite setting contrarily to min-based DL-lite. An encoding of the inconsistency degree computing problem as a WM2HSAT problem has been proposed. This transformation was then used to propose an algorithm using a WM2HSAT solver to compute the inconsistency degree. As a future work, we plan to generalize our framework to arbitrary DL-Lite bases where the TBox axioms are not fully certain.

**Acknowledgments** This work has been supported from the european project H2020 Marie Sklodowska-Curie Actions (MSCA) research and Innovation Staff Exchange (RISE): AniAge (High Dimensional Heterogeneous Data based Animation Techniques for Southeast Asian Intangible Cultural Heritage. This work has also been supported by the French National Research Agency ASPIQ project ANR-12-BS02-0003.

## References

1. Artale A, Calvanese DR, Kontchakov Zakharyaschev M (2009) The dl-lite family and relations. J Artif Intell Res 36:1–69
2. Baader F, McGuinness L, Nardi D, Patel-Schneider P (2003) The description logic handbook: theory, implementation, and applications. Cambridge University Press
3. Benferhat S, Bouraoui Z (2015) Min-based possibilistic dl-lite. J Logic Comput 2015. doi:10.1093/logcom/exv014, first published online April 12
4. Bobillo F, Delgado M, Gmez-Romero J (2012) Delorean: a reasoner for fuzzy owl 2. Expert Syst Appl 39(1):258–272
5. Calvanese D, De Giacomo G, Lembo D, Lenzerini M, Rosati R (2007) Tractable reasoning and efficient query answering in description logics: the dl-lite family. J Autom Reason 39(3):385–429
6. Dubois D, Mengin J, Prade H (2006) Possibilistic uncertainty and fuzzy features in description logic: A preliminary discussion. In: Capturing Intelligence: Fuzzy Logic and the Semantic WEB, Elsevier, pp 101–113
7. Lukasiewicz T, Straccia U (2009) Description logic programs under probabilistic uncertainty and fuzzy vagueness. Int J Approx Reason 50(6):837–853
8. Pan J, Stamou G, Stoilos G, Thomas E (2007) Expressive querying over fuzzy dl-lite ontologies. In: Proceedings of the 2007 International Workshop on Description Logics
9. Qi G, Ji Q, Pan J, Du J (2011) Extending description logics with uncertainty reasoning in possibilistic logic. Int J Intell Syst 26(4):353–381
10. Straccia U (2013) Foundations of fuzzy logic and semantic web languages. CRC Press

# Asymptotics of Predictive Distributions

**Patrizia Berti, Luca Pratelli and Pietro Rigo**

**Abstract** Let $(X_n)$ be a sequence of random variables, adapted to a filtration $(\mathcal{G}_n)$, and let $\mu_n = (1/n)\,\sum_{i=1}^n \delta_{X_i}$ and $a_n(\cdot) = P(X_{n+1} \in \cdot \mid \mathcal{G}_n)$ be the empirical and the predictive measures. We focus on $\|\mu_n - a_n\| = \sup_{B\in\mathcal{D}} |\mu_n(B) - a_n(B)|$, where $\mathcal{D}$ is a class of measurable sets. Conditions for $\|\mu_n - a_n\| \to 0$, almost surely or in probability, are given. Also, to determine the rate of convergence, the asymptotic behavior of $r_n\,\|\mu_n - a_n\|$ is investigated for suitable constants $r_n$. Special attention is paid to $r_n = \sqrt{n}$. The sequence $(X_n)$ is exchangeable or, more generally, conditionally identically distributed.

## 1 Introduction

### 1.1 *The Problem*

Throughout, $S$ is a Polish space and $X = (X_n : n \geq 1)$ a sequence of $S$-valued random variables on the probability space $(\Omega, \mathcal{A}, P)$. Further, $\mathcal{B}$ is the Borel $\sigma$-field on $S$ and $\mathcal{G} = (\mathcal{G}_n : n \geq 0)$ a filtration on $(\Omega, \mathcal{A}, P)$. We fix a subclass $\mathcal{D} \subset \mathcal{B}$ and we let $\|\cdot\|$ denote the sup-norm over $\mathcal{D}$, namely, $\|\alpha - \beta\| = \sup_{B\in\mathcal{D}} |\alpha(B) - \beta(B)|$ whenever $\alpha$ and $\beta$ are probabilities on $\mathcal{B}$.

Let

$$\mu_n = (1/n)\sum_{i=1}^n \delta_{X_i} \quad \text{and} \qquad a_n(\cdot) = P(X_{n+1} \in \cdot \mid \mathcal{G}_n).$$

P. Berti
Universita' di Modena e Reggio-Emilia, Modena, Italy
e-mail: patrizia.berti@unimore.it

L. Pratelli
Accademia Navale di Livorno, Livorno, Italy
e-mail: pratel@mail.dm.unipi.it

P. Rigo (✉)
Universita' di Pavia, Pavia, Italy
e-mail: pietro.rigo@unipv.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_7

Both $\mu_n$ and $a_n$ are regarded as random probability measures on $\mathcal{B}$; $\mu_n$ is the empirical measure and (if $X$ is $\mathcal{G}$-adapted) $a_n$ is the predictive measure.

Under some conditions, $\mu_n(B) - a_n(B) \xrightarrow{a.s.} 0$ for fixed $B \in \mathcal{B}$. In that case, a (natural) question is whether $\mathcal{D}$ is such that $\|\mu_n - a_n\| \xrightarrow{a.s.} 0$.

Such question is addressed in this paper. Conditions for $\|\mu_n - a_n\| \to 0$, almost surely or in probability, are given. Also, to determine the rate of convergence, the asymptotic behavior of $r_n \|\mu_n - a_n\|$ is investigated for suitable constants $r_n$. Special attention is paid to $r_n = \sqrt{n}$. The sequence $X$ is assumed to be exchangeable or, more generally, conditionally identically distributed (see Sect. 2).

Our main concern is to connect and unify a few results from [1–4]. Thus, this paper is essentially a survey. However, in addition to report known facts, some new results and examples are given. This is actually the case of Theorem 1(d), Corollary 1 and Examples 1–3.

## 1.2 Heuristics

There are various (non-independent) reasons for investigating $\mu_n - a_n$. We now list a few of them under the assumption that $\mathcal{G} = \mathcal{G}^X$, where $\mathcal{G}_0^X = \{\emptyset, \Omega\}$ and $\mathcal{G}_n^X = \sigma(X_1, \dots, X_n)$. Most remarks, however, apply to any filtration $\mathcal{G}$ which makes $X$ adapted.

- **Empirical processes for non-ergodic data**. Slightly abusing terminology, say that $X$ is ergodic if $P$ is 0–1 valued on the sub-$\sigma$-field $\sigma\big(\limsup_n \mu_n(B) : B \in \mathcal{B}\big)$. In real problems, $X$ is often non-ergodic. Most stationary sequences, for instance, fail to be ergodic. Or else, an exchangeable sequence is ergodic if and only if is i.i.d. Now, if $X$ is i.i.d., the empirical process is defined as $G_n = \sqrt{n}\,(\mu_n - \mu_0)$ where $\mu_0$ is the probability distribution of $X_1$. But this definition has various drawbacks when $X$ is not ergodic; see [5]. In fact, unless $X$ is i.i.d., the probability distribution of $X$ is not determined by that of $X_1$. More importantly, if $G_n$ converges in distribution in $l^\infty(\mathcal{D})$ (the metric space $l^\infty(\mathcal{D})$ is recalled before Corollary 1) then $\|\mu_n - \mu_0\| = n^{-1/2}\|G_n\| \xrightarrow{P} 0$. But $\|\mu_n - \mu_0\|$ typically fails to converge to 0 in probability when $X$ is not ergodic. Thus, empirical processes for non-ergodic data should be defined in some different way. In this framework, a meaningful option is to replace $\mu_0$ with $a_n$, namely, to let $G_n = \sqrt{n}\,(\mu_n - a_n)$.
- **Bayesian predictive inference**. In a number of problems, the main goal is to evaluate $a_n$ but the latter can not be obtained in closed form. Thus, $a_n$ is to be estimated by the available data. Under some assumptions, a reasonable estimate of $a_n$ is just $\mu_n$. In these situations, the asymptotic behavior of the error $\mu_n - a_n$ plays a role. For instance, $\mu_n$ is a consistent estimate of $a_n$ provided $\|\mu_n - a_n\| \longrightarrow 0$ in some sense.

- **Predictive distributions of exchangeable sequences**. Let $X$ be exchangeable. Just very little is known on the general form of $a_n$ for given $n$, and a representation theorem for $a_n$ would be actually a major breakthrough. Failing the latter, to fix the asymptotic behavior of $\mu_n - a_n$ contributes to fill the gap.
- **de Finetti**. Historically, one reason for introducing exchangeability (possibly, the main reason) was to justify observed frequencies as predictors of future events. See [8–10]. In this sense, to focus on $\mu_n - a_n$ is in line with de Finetti's ideas. Roughly speaking, $\mu_n$ should be a good substitute of $a_n$ in the exchangeable case.

## 2 Conditionally Identically Distributed Sequences

The sequence $X$ is *conditionally identically distributed* (c.i.d.) with respect to $\mathcal{G}$ if it is $\mathcal{G}$-adapted and $P\big(X_k \in \cdot \mid \mathcal{G}_n\big) = P\big(X_{n+1} \in \cdot \mid \mathcal{G}_n\big)$ a.s. for all $k > n \geq 0$. Roughly speaking, at each time $n \geq 0$, the future observations $(X_k : k > n)$ are identically distributed given the past $\mathcal{G}_n$. When $\mathcal{G} = \mathcal{G}^X$, the filtration $\mathcal{G}$ is not mentioned at all and $X$ is just called c.i.d. Then, $X$ is c.i.d. if and only if $\big(X_1, \ldots, X_n, X_{n+2}\big) \sim \big(X_1, \ldots, X_n, X_{n+1}\big)$ for all $n \geq 0$.

Exchangeable sequences are c.i.d. while the converse is not true. Indeed, $X$ is exchangeable if and only if it is stationary and c.i.d. We refer to [3] for more on c.i.d. sequences. Here, it suffices to mention a last fact.

If $X$ is c.i.d., there is a random probability measure $\mu$ on $\mathcal{B}$ such that $\mu_n(B) \xrightarrow{a.s.} \mu(B)$ for every $B \in \mathcal{B}$. As a consequence, if $X$ is c.i.d. with respect to $\mathcal{G}$, for each $n \geq 0$ and $B \in \mathcal{B}$ one obtains

$$\begin{aligned} E\big\{\mu(B) \mid \mathcal{G}_n\big\} &= \lim_m E\big\{\mu_m(B) \mid \mathcal{G}_n\big\} = \lim_m \frac{1}{m} \sum_{k=n+1}^{m} P\big(X_k \in B \mid \mathcal{G}_n\big) \\ &= P\big(X_{n+1} \in B \mid \mathcal{G}_n\big) = a_n(B) \quad \text{a.s.} \end{aligned}$$

In particular, $a_n(B) = E\big\{\mu(B) \mid \mathcal{G}_n\big\} \xrightarrow{a.s.} \mu(B)$ and $\mu_n(B) - a_n(B) \xrightarrow{a.s.} 0$.

From now on, $X$ is c.i.d. with respect to $\mathcal{G}$. In particular, $X$ is identically distributed and $\mu_0$ denotes the probability distribution of $X_1$. We also let

$$W_n = \sqrt{n}\,(\mu_n - \mu),$$

where $\mu$ is the random probability measure on $\mathcal{B}$ introduced above. Note that, if $X$ is i.i.d., then $\mu = \mu_0$ a.s. and $W_n$ reduces to the usual empirical process.

# 3 Results

Let $\mathcal{D} \subset \mathcal{B}$. To avoid measurability problems, $\mathcal{D}$ is assumed to be *countably determined*. This means that there is a countable subclass $\mathcal{D}_0 \subset \mathcal{D}$ such that $\|\alpha - \beta\| = \sup_{B \in \mathcal{D}_0} |\alpha(B) - \beta(B)|$ for all probabilities $\alpha$, $\beta$ on $\mathcal{B}$. For instance, $\mathcal{D} = \mathcal{B}$ is countably determined (for $\mathcal{B}$ is countably generated). Or else, if $S = \mathbb{R}^k$, then $\mathcal{D} = \{(-\infty, t] : t \in \mathbb{R}^k\}$, $\mathcal{D} = \{\text{closed balls}\}$ and $\mathcal{D} = \{\text{closed convex sets}\}$ are countably determined.

## 3.1 A General Criterion

Since $a_n(B) = E\big\{\mu(B) \mid \mathcal{G}_n\big\}$ a.s. for each $B \in \mathcal{B}$ and $\mathcal{D}$ is countably determined, one obtains

$$\|\mu_n - a_n\| = \sup_{B \in \mathcal{D}_0} |E\big\{\mu_n(B) - \mu(B) \mid \mathcal{G}_n\big\}| \leq E\big\{\|\mu_n - \mu\| \mid \mathcal{G}_n\big\} \quad \text{a.s.}$$

This simple inequality has some nice consequences. Recall that $\mathcal{D}$ is a *universal Glivenko-Cantelli class* if $\|\mu_n - \mu_0\| \xrightarrow{a.s.} 0$ whenever $X$ is i.i.d.

**Theorem 1** *Suppose $\mathcal{D}$ is countably determined and $X$ is c.i.d. with respect to $\mathcal{G}$. Then,*

*(a)* $\|\mu_n - a_n\| \xrightarrow{a.s.} 0$ *if* $\|\mu_n - \mu\| \xrightarrow{a.s.} 0$ *and* $\|\mu_n - a_n\| \xrightarrow{P} 0$ *if* $\|\mu_n - \mu\| \xrightarrow{P} 0$.
*(b)* $\|\mu_n - a_n\| \xrightarrow{a.s.} 0$ *provided $X$ is exchangeable, $\mathcal{G} = \mathcal{G}^X$ and $\mathcal{D}$ is a universal Glivenko-Cantelli class.*
*(c)* $r_n \|\mu_n - a_n\| \xrightarrow{P} 0$ *whenever the constants $r_n$ satisfy $r_n/\sqrt{n} \to 0$ and* $\sup_n E\big\{\|W_n\|^b\big\} < \infty$ *for some $b \geq 1$.*
*(d)* $n^u \|\mu_n - a_n\| \xrightarrow{a.s.} 0$ *whenever $u < 1/2$ and* $\sup_n E\big\{\|W_n\|^b\big\} < \infty$ *for each $b \geq 1$.*

*Proof* Since $\|\mu_n - \mu\| \leq 1$, point (a) follows from the martingale convergence theorem in the version of [7]. (If $\|\mu_n - \mu\| \xrightarrow{P} 0$, it suffices to apply an obvious argument based on subsequences). Next, suppose $X$, $\mathcal{G}$ and $\mathcal{D}$ are as in (b). By de Finetti's theorem, conditionally on $\mu$, the sequence $X$ is i.i.d. with common distribution $\mu$. Since $\mathcal{D}$ is a universal Glivenko-Cantelli class, it follows that $P\big(\|\mu_n - \mu\| \to 0\big) = \int P\big\{\|\mu_n - \mu\| \to 0 \mid \mu\big\}\, dP = \int 1 dP = 1$. Hence, (b) is a consequence of (a). As to (c), just note that

$$E\Big\{\big(r_n \, \|\mu_n - a_n\|\big)^b\Big\} \leq r_n^b \, E\big\{\|\mu_n - \mu\|^b\big\} = (r_n/\sqrt{n})^b \, E\big\{\|W_n\|^b\big\}.$$

Finally, as to (d), fix $u < 1/2$ and take $b$ such that $b(1/2 - u) > 1$. Then,

$$\sum_n P\big(n^u \|\mu_n - a_n\| > \epsilon\big) \leq \sum_n \frac{E\big\{\|\mu_n - a_n\|^b\big\}}{\epsilon^b\, n^{-ub}} \leq \sum_n \frac{E\big\{\|\mu_n - \mu\|^b\big\}}{\epsilon^b\, n^{-ub}}$$

$$= \sum_n \frac{E\big\{\|W_n\|^b\big\}}{\epsilon^b\, n^{(1/2-u)b}} \leq \sum_n \frac{\text{const}}{n^{(1/2-u)b}} < \infty \quad \text{for each } \epsilon > 0.$$

Therefore, $n^u \|\mu_n - a_n\| \xrightarrow{a.s.} 0$ because of the Borel-Cantelli lemma.

Some remarks are in order.

Theorem 1 is essentially known. Apart from (d), it is implicit in [2, 4].

If $X$ is exchangeable, the second part of (a) is redundant. In fact, $\|\mu_n - \mu_0\|$ converges a.s. (not necessarily to 0) whenever $X$ is i.i.d. Applying de Finetti's theorem as in the proof of Theorem 1(b), it follows that $\|\mu_n - \mu\|$ converges a.s. even if $X$ is exchangeable. Thus, $\|\mu_n - \mu\| \xrightarrow{P} 0$ implies $\|\mu_n - \mu\| \xrightarrow{a.s.} 0$.

Sometimes, the condition in (a) is necessary as well, namely, $\|\mu_n - a_n\| \xrightarrow{a.s.} 0$ if and only if $\|\mu_n - \mu\| \xrightarrow{a.s.} 0$. For instance, this happens when $\mathcal{G} = \mathcal{G}^X$ and $\mu \ll \lambda$ a.s., where $\lambda$ is a (non-random) $\sigma$-finite measure on $\mathcal{B}$. In this case, in fact, $\|a_n - \mu\| \xrightarrow{a.s.} 0$ by [6, Theorem 1].

Several examples of universal Glivenko-Cantelli classes are available; see [11] and references therein. Similarly, for many choices of $\mathcal{D}$ and $b \geq 1$ there is a universal constant $c(b)$ such that $\sup_n E\big\{\|W_n\|^b\big\} \leq c(b)$ provided $X$ is i.i.d.; see e.g. [11, Sects. 2.14.1 and 2.14.2]. In these cases, de Finetti's theorem yields $\sup_n E\big\{\|W_n\|^b\big\} \leq c(b)$ even if $X$ is exchangeable. Thus, points (b)–(d) are especially useful when $X$ is exchangeable.

In (c), convergence in probability can not be replaced by a.s. convergence. As a trivial example, take $\mathcal{D} = \mathcal{B}, \mathcal{G} = \mathcal{G}^X, r_n = \sqrt{\frac{n}{\log\log n}}$, and $X$ an i.i.d. sequence of indicators. Letting $p = P(X_1 = 1)$, one obtains $E\big\{\|W_n\|^2\big\} = n\, E\big\{\big(\mu_n\{1\} - p\big)^2\big\} = p\,(1-p)$ for all $n$. However, the LIL yields

$$\limsup_n r_n \|\mu_n - a_n\| = \limsup_n \frac{|\sum_{i=1}^n (X_i - p)|}{\sqrt{n}\, \log\log n} = \sqrt{2\,p\,(1-p)} \quad \text{a.s.}$$

We finally give a couple of examples.

*Example 1* Let $\mathcal{D} = \mathcal{B}$. If $X$ is i.i.d., then $\|\mu_n - \mu_0\| \xrightarrow{a.s.} 0$ if and only if $\mu_0$ is discrete. By de Finetti's theorem, it follows that $\|\mu_n - \mu\| \xrightarrow{a.s.} 0$ whenever $X$ is exchangeable and $\mu$ is a.s. discrete. Thus, under such assumptions and $\mathcal{G} = \mathcal{G}^X$, Theorem 1(a) implies $\|\mu_n - a_n\| \xrightarrow{a.s.} 0$. This result has possible practical interest. In fact, in Bayesian nonparametrics, most priors are such that $\mu$ is a.s. discrete.

*Example 2* Let $S = \mathbb{R}^k$ and $\mathcal{D} = \{\text{closed convex sets}\}$. Given any probability $\alpha$ on $\mathcal{B}$, denote by $\alpha^{(c)} = \alpha - \sum_x \alpha\{x\}\delta_x$ the continuous part of $\alpha$. If $X$ is i.i.d. and $\mu_0^{(c)} \ll m$, where $m$ is Lebesgue measure, then $\|\mu_n - \mu_0\| \xrightarrow{a.s.} 0$. Applying Theorem 1(a) again, one obtains $\|\mu_n - a_n\| \xrightarrow{a.s.} 0$ provided $X$ is exchangeable, $\mathcal{G} = \mathcal{G}^X$ and $\mu^{(c)} \ll m$ a.s. While "morally true", this argument does not work for $\mathcal{D} = \{\text{Borel convex sets}\}$ since the latter choice of $\mathcal{D}$ is not countably determined.

## 3.2 The Dominated Case

In this Subsection, $\mathcal{G} = \mathcal{G}^X$, $\mathcal{A} = \sigma\big(\cup_n \mathcal{G}_n^X\big)$, $Q$ is a probability on $(\Omega, \mathcal{A})$ and $b_n(\cdot) = Q(X_{n+1} \in \cdot \mid \mathcal{G}_n)$ is the predictive measure under $Q$. Also, we say that $Q$ is a Ferguson-Dirichlet law if

$$b_n(\cdot) = \frac{c\,Q(X_1 \in \cdot) + n\,\mu_n(\cdot)}{c+n}, \quad Q\text{-a.s. for some constant } c > 0.$$

If $P \ll Q$, the asymptotic behavior of $\mu_n - a_n$ under $P$ should be affected by that of $\mu_n - b_n$ under $Q$. This (rough) idea is realized by the next result.

**Theorem 2** (Theorems 1 and 2 of [4]) *Suppose $\mathcal{D}$ is countably determined, $X$ is c.i.d., and $P \ll Q$. Then, $\sqrt{n}\,\|\mu_n - a_n\| \xrightarrow{P} 0$ provided $\sqrt{n}\,\|\mu_n - b_n\| \xrightarrow{Q} 0$ and the sequence $(W_n)$ is uniformly integrable under both $P$ and $Q$. In addition, $n\,\|\mu_n - a_n\|$ converges a.s. to a finite limit whenever $Q$ is a Ferguson-Dirichlet law,* $\sup_n E_Q\big\{\|W_n\|^2\big\} < \infty$, *and*

$$\sup_n n\,\Big\{E_Q\big\{(dP/dQ)^2\big\} - E_Q\big\{E_Q(dP/dQ \mid \mathcal{G}_n)^2\big\}\Big\} < \infty.$$

To make Theorem 2 effective, the condition $P \ll Q$ should be given a simple characterization. This happens in at least one case.

Let $S$ be finite, say $S = \{x_1, \ldots, x_k, x_{k+1}\}$, $X$ exchangeable and $\mu_0\{x\} > 0$ for all $x \in S$. Then $P \ll Q$, with $Q$ a Ferguson-Dirichlet law, if and only if the distribution of $\big(\mu\{x_1\}, \ldots, \mu\{x_k\}\big)$ is absolutely continuous (with respect to Lebesgue measure). This fact is behind the next result.

**Theorem 3** (Corollaries 4 and 5 of [4]) *Suppose $S = \{0, 1\}$ and $X$ is exchangeable. Then, $\sqrt{n}\,\big(\mu_n\{1\} - a_n\{1\}\big) \xrightarrow{P} 0$ whenever the distribution of $\mu\{1\}$ is absolutely continuous. Moreover, $n\,\big(\mu_n\{1\} - a_n\{1\}\big)$ converges a.s. (to a finite limit) provided the distribution of $\mu\{1\}$ is absolutely continuous with an almost Lipschitz density.*

In Theorem 3, a real function $f$ on $(0, 1)$ is said to be *almost Lipschitz* in case $x \mapsto f(x)x^u(1-x)^v$ is Lipschitz on $(0, 1)$ for some reals $u,\ v < 1$.

A consequence of Theorem 3 is to be stressed. For each $B \in \mathcal{B}$, define

$$T_n(B) = \sqrt{n}\,\Big\{a_n(B) - P\big\{X_{n+1} \in B \mid \mathcal{G}_n^B\big\}\Big\}$$

where $\mathcal{G}_n^B = \sigma\big(I_B(X_1), \dots, I_B(X_n)\big)$. Also, let $l^\infty(\mathcal{D})$ be the set of real bounded functions on $\mathcal{D}$, equipped with uniform distance. In the next result, $W_n$ is regarded as a random element of $l^\infty(\mathcal{D})$ and convergence in distribution is meant in Hoffmann-Jørgensen's sense; see [11].

**Corollary 1** *Let $\mathcal{D}$ be countably determined and $X$ exchangeable. Suppose*

*(i)* *$\mu(B)$ has an absolutely continuous distribution for each $B \in \mathcal{D}$ such that* $0 < P(X_1 \in B) < 1$*;*
*(ii)* *the sequence $(\|W_n\|)$ is uniformly integrable;*
*(iii)* *$W_n$ converges in distribution to a tight limit in $l^\infty(\mathcal{D})$.*

*Then, $\sqrt{n}\,\|\mu_n - a_n\| \xrightarrow{P} 0$ if and only if $T_n(B) \xrightarrow{P} 0$ for each $B \in \mathcal{D}$.*

*Proof* Let $U_n(B) = \sqrt{n}\,\Big\{\mu_n(B) - P\big\{X_{n+1} \in B \mid \mathcal{G}_n^B\big\}\Big\}$. Then, $U_n(B) \xrightarrow{P} 0$ for each $B \in \mathcal{D}$. In fact, $U_n(B) = 0$ a.s. if $P(X_1 \in B) \in \{0, 1\}$. Otherwise, $U_n(B) \xrightarrow{P} 0$ follows from Theorem 3, since $(I_B(X_n))$ is an exchangeable sequence of indicators and $\mu(B)$ has an absolutely continuous distribution. Next, suppose $T_n(B) \xrightarrow{P} 0$ for each $B \in \mathcal{D}$. Letting $C_n = \sqrt{n}\,(\mu_n - a_n)$, we have to prove that $\|C_n\| \xrightarrow{P} 0$. Equivalently, regarding $C_n$ as a random element of $l^\infty(\mathcal{D})$, we have to prove that $C_n(B) \xrightarrow{P} 0$ for fixed $B \in \mathcal{D}$ and the sequence $(C_n)$ is asymptotically tight; see e.g. [11, Sect. 1.5]. Given $B \in \mathcal{D}$, since both $U_n(B)$ and $T_n(B)$ converge to 0 in probability, then $C_n(B) = U_n(B) - T_n(B) \xrightarrow{P} 0$. Moreover, since $C_n(B) = E\big\{W_n(B) \mid \mathcal{G}_n\big\}$ a.s., the asymptotic tightness of $(C_n)$ follows from (ii) and (iii); see [3, Remark 4.4]. Hence, $\|C_n\| \xrightarrow{P} 0$. Conversely, if $\|C_n\| \xrightarrow{P} 0$, one trivially obtains

$$|T_n(B)| = |U_n(B) - C_n(B)| \leq |U_n(B)| + \|C_n\| \xrightarrow{P} 0 \quad \text{for each } B \in \mathcal{D}.$$

If $X$ is exchangeable, it frequently happens that $\sup_n E\big\{\|W_n\|^2\big\} < \infty$, which in turn implies condition (ii). Similarly, (iii) is not unusual. As an example, conditions (ii) and (iii) hold if $S = \mathbb{R}$, $\mathcal{D} = \{(-\infty, t] : t \in \mathbb{R}\}$ and $\mu_0$ is discrete or $P(X_1 = X_2) = 0$; see [3, Theorem 4.5].

Unfortunately, as shown by the next example, $T_n(B)$ may fail to converge to 0 even if $\mu(B)$ has an absolutely continuous distribution. This suggests the following general question. In the exchangeable case, in addition to $\mu_n(B)$, which further information is required to evaluate $a_n(B)$? Or at least, are there reasonable conditions for $T_n(B) \xrightarrow{P} 0$? Even if intriguing, to our knowledge, such a question does not have a satisfactory answer.

*Example 3* Let $S = \mathbb{R}$ and $X_n = Y_n\, Z^{-1}$, where $Y_n$ and $Z$ are independent real random variables, $Y_n \sim N(0, 1)$ for all $n$, and $Z$ has an absolutely continuous distribution supported by $[1, \infty)$. Conditionally on $Z$, the sequence $X = (X_1, X_2, \dots)$ is i.i.d. with common distribution $N(0, Z^{-2})$. Thus, $X$ is exchangeable and $\mu(B) = P(X_1 \in B \mid Z) = f_B(Z)$ a.s., where

$$f_B(z) = (2\,\pi)^{-1/2} z \int_B \exp\big(-(xz)^2/2\big)\,dx \quad \text{for } B \in \mathcal{B} \text{ and } z \geq 1.$$

Fix $B \in \mathcal{B}$, with $B \subset [1, \infty)$ and $P(X_1 \in B) > 0$, and define $C = \{-x : x \in B\}$. Since $f_B = f_C$, then $\mu(B) = \mu(C)$ a.s. Further, $\mu(B)$ has an absolutely continuous distribution, for $f_B$ is differentiable and $f_B' \neq 0$. Nevertheless, one between $T_n(B)$ and $T_n(C)$ does not converge to 0in probability. Define in fact $g = I_B - I_C$ and $R_n = n^{-1/2} \sum_{i=1}^{n} g(X_i)$. Since $\mu(g) = \mu(B) - \mu(C) = 0$ a.s., then $R_n$ converges stably to the kernel $N(0,\, 2\mu(B))$; see [3, Theorem 3.1]. On the other hand, since $E\big\{g(X_{n+1}) \mid \mathcal{G}_n\big\} = E\big\{\mu(g) \mid \mathcal{G}_n\big\} = 0$ a.s., one obtains

$$R_n = \sqrt{n}\,\big\{\mu_n(B) - \mu_n(C)\big\} = T_n(C) - T_n(B) +$$
$$+ \sqrt{n}\,\Big\{\mu_n(B) - P\big\{X_{n+1} \in B \mid \mathcal{G}_n^B\big\}\Big\} - \sqrt{n}\,\Big\{\mu_n(C) - P\big\{X_{n+1} \in C \mid \mathcal{G}_n^C\big\}\Big\}.$$

Hence, if $T_n(B) \xrightarrow{P} 0$ and $T_n(C) \xrightarrow{P} 0$, Corollary 1 (applied with $\mathcal{D} = \{B, C\}$) implies the contradiction $R_n \xrightarrow{P} 0$.

## References

1. Berti P, Rigo P (1997) A Glivenko-Cantelli theorem for exchangeable random variables. Stat Probab Lett 32:385–391
2. Berti P, Mattei A, Rigo P (2002) Uniform convergence of empirical and predictive measures. Atti Sem Mat Fis Univ Modena 50:465–477
3. Berti P, Pratelli L, Rigo P (2004) Limit theorems for a class of identically distributed random variables. Ann Probab 32:2029–2052
4. Berti P, Crimaldi I, Pratelli L, Rigo P (2009) Rate of convergence of predictive distributions for dependent data. Bernoulli 15:1351–1367
5. Berti P, Pratelli L, Rigo P (2012) Limit theorems for empirical processes based on dependent data. Electron J Probab 17:1–18
6. Berti P, Pratelli L, Rigo P (2013) Exchangeable sequences driven by an absolutely continuous random measure. Ann Probab 41:2090–2102
7. Blackwell D, Dubins LE (1962) Merging of opinions with increasing information. Ann Math Stat 33:882–886
8. Cifarelli DM, Regazzini E (1996) De Finetti's contribution to probability and statistics. Stat Sci 11:253–282
9. Cifarelli DM, Dolera E, Regazzini E (2016) Frequentistic approximations to Bayesian prevision of exchangeable random elements. arXiv:1602.01269v1
10. Fortini S, Ladelli L, Regazzini E (2000) Exchangeability, predictive distributions and parametric models. Sankhya A 62:86–109
11. van der Vaart A, Wellner JA (1996) Weak convergence and empirical processes. Springer

# Independent $k$-Sample Equality Distribution Test Based on the Fuzzy Representation

**Angela Blanco-Fernández and Ana B. Ramos-Guajardo**

**Abstract** Classical tests for the equality of distributions of real-valued random variables are widely applied in Statistics. When the normality assumption for the variables fails, non-parametric techniques are to be considered; Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman tests, among other alternatives. Fuzzy representations of real-valued random variables have been recently shown to describe in an effective way the statistical behaviour of the variables. Indeed, the expected value of certain fuzzy representations fully characterizes the distribution of the variable. The aim of this paper is to use this characterization to test the equality of distribution for two or more real-valued random variables, as an alternative to classical procedures. The inferential problem is solved through a parametric test for the equality of expectations of fuzzy-valued random variables. Theoretical results on inferences for fuzzy random variables support the validity of the test. Besides, simulation studies and practical applications show the empirical goodness of the method.

## 1 Introduction

The development of statistical methods for fuzzy random variables has increased exponentially in last decades, from the seminal ideas on fuzziness by Zadeh [16]. In some situations, experimental data are not precise observations, represented by fixed categories or point-valued real numbers, and modelled by real-valued variables. The outcomes of the experiment might be more imprecise or *fuzzy*, in the sense that they are not represented by just a point value, but a set of values, an interval, or even a function. Imprecise experimental data can be effectively modelled by means of

A. Blanco-Fernández (✉) · A.B. Ramos-Guajardo
Departament of Statistics and Operational Research,
University of Oviedo, C/Calvo Sotelo, s/n, 33007 Oviedo, Spain
e-mail: blancoangela@uniovi.es

A.B. Ramos-Guajardo
e-mail: ramosana@uniovi.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_8

fuzzy-valued variables. Additionally to the imprecision on the data, the randomness on the data generation process drives to the formalization of fuzzy-valued random variables (FRVs).

Powerful exploratory, probabilistic and inferential studies for fuzzy random variables have been deeply investigated in the literature. It is important to remark that fuzzy data can be seen under two different perspectives, usually called *ontic* and *epistemic* views of fuzzy data, and the statistical treatment of the variables in each line is radically different. In few words, the epistemic approach considers the fuzzy data as imprecise observations or descriptions of crisp (but unknown) quantities. Statistical methods are focused to draw conclusions for the original real-valued variable, and they generally transfer the imprecision to methods and results [4, 5, 10, 12]. Alternatively, ontic fuzzy data are treated as precise entities representing the outcomes of the experiment, belonging to the corresponding space of functions instead of the space of real numbers. In this case, statistical methods try to mimic classical techniques to draw conclusions directly to the fuzzy-valued variables modelling the experiment [2, 6, 7, 11, 13]. Further discussions on the two frameworks can be found in [2, 5].

Besides their own statistical analysis, fuzzy-valued random variables in the ontic perspective have been also shown as a powerful tool to obtain statistical conclusions to classical real-valued random variables [1, 3, 8, 9]. Exploratory and inferential studies for real random variables have been developed through the so-called fuzzy representation of the variable, defined, roughly speaking, by applying a fuzzy operator to the original variable and *fuzzifying* its values. The key idea is that it is not included imprecision in the data gratuitously, but this transformation is very effective to certain statistical purposes. The aim of this paper is to extend this line of research to test the equality of two or more real-valued distributions based on the fuzzy representation of the variables. The rest of the paper is organized as follows. In Sect. 2, the main concepts concerning fuzzy random variables and the concept of fuzzy representation of a real-valued variable are recalled. The inferential studies on the equality of real-valued distributions are presented in Sect. 3. Theoretical and empirical results on the proposed tests are shown. Finally, some conclusions and future problems are commented in Sect. 4.

## 2 Preliminaries

Let $\mathcal{F}_c(\mathbb{R})$ denote the class of fuzzy sets $U : \mathbb{R} \to [0, 1]$ such that $U_\alpha \in \mathcal{K}_c(\mathbb{R})$ for all $\alpha \in [0, 1]$, where $\mathcal{K}_c(\mathbb{R})$ is the family of all non-empty closed and bounded intervals of $\mathbb{R}$, the $\alpha$-levels of $U$ are defined as $U_\alpha = \{x \in \mathbb{R} | U(x) \geq \alpha\}$ if $\alpha \in (0, 1]$, and $U_0$ is the closure of the support of $U$.

The usual arithmetic between fuzzy sets is based on Zadeh's extension principle [16]. It agrees levelwise with the Minkowski addition and the product by scalars for intervals. Given $U, V \in \mathcal{F}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$, $U + V$ and $\lambda U$ are defined such that $(U + V)_\alpha = U_\alpha + V_\alpha = \{u + v : u \in U_\alpha, v \in V_\alpha\}$ and $(\lambda U)_\alpha = \lambda U_\alpha = \{\lambda u : u \in U_\alpha\}$, for all $\alpha \in [0, 1]$.

The space $\mathcal{F}_c(\mathbb{R})$ can be embedded into a convex and closed cone of $\mathcal{L}^2(\{-1, 1\} \times [0, 1])$ by means of the *support function* [11], defined for any $U \in \mathcal{F}_c(\mathbb{R})$ as $s_U : \{-1, 1\} \times [0, 1] \to \mathbb{R}$ such that $s_U(u, \alpha) = \sup_{v \in U_\alpha} \langle u, v \rangle$. It is important to note that, although this embedding permits good operational properties, the statistical processing of fuzzy sets cannot be directly transferred to $\mathcal{L}^2(\{-1, 1\} \times [0, 1])$; it must always be guaranteed that the results remain coherently into the cone.

In order to measure distances between fuzzy sets, the family of metrics $D_\theta^\varphi$ in $\mathcal{F}_c(\mathbb{R})$ [14] is defined as

$$D_\theta^\varphi(U, V) = \sqrt{\int_{(0,1]} \Big((\mathrm{mid}U_\alpha - \mathrm{mid}V_\alpha)^2 + \theta(\mathrm{spr}U_\alpha - \mathrm{spr}V_\alpha)^2\Big) d\varphi(\alpha)},$$

with $\theta > 0$, $\varphi$ is associated with a bounded density measure with positive mass in (0, 1], and $\mathrm{mid}U_\alpha/\mathrm{spr}U_\alpha$ are the mid-point/radius of the interval $U_\alpha \in \mathcal{K}_c(\mathbb{R})$, respectively, i.e. $U_\alpha = [\mathrm{mid}U_\alpha \pm \mathrm{spr}U_\alpha]$ for all $\alpha \in [0, 1]$.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. A mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ is a random fuzzy set (RFS) (or random fuzzy variable) if it is Borel-measurable with respect to $\mathcal{B}_{D_\theta^\varphi}$, the $\sigma$-field generated by the topology induced by the metric $D_\theta^\varphi$ on the space $\mathcal{F}_c(\mathbb{R})$.

The central tendency of a RFS is usually measured by the *Aumann expectation of* $\mathcal{X}$. If $\max\{\|\inf \mathcal{X}_0\|, \|\sup \mathcal{X}_0\|\} \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$, it is defined as the unique fuzzy set $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$ such that

$$\big(\widetilde{E}(\mathcal{X})\big)_\alpha = \text{ Kudo-Aumann's integral of } \mathcal{X}_\alpha = [E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)],$$

for all $\alpha \in [0, 1]$. Given $\{\mathcal{X}_i\}_{i=1}^n$ a simple random sample of size $n$ from $\mathcal{X}$,the associated sample mean is defined as $\overline{\mathcal{X}_n} = (1/n) \sum_{i=1}^n \mathcal{X}_i$.

Let $\gamma : \mathbb{R} \to \mathcal{F}_c(\mathbb{R})$ the mapping transforming each $x \in \mathbb{R}$ into the fuzzy set $\gamma(x)$ whose $\alpha$-levels are given by

$$\Big(\gamma(x)\Big)_\alpha = \Big[f_L(x) - g_L(x)(1-\alpha)^{1/h_L(x)}, f_R(x) + g_R(x)(1-\alpha)^{1/h_R(x)}\Big],$$

for all $\alpha \in [0, 1]$, where $f_L, f_R : \mathbb{R} \to \mathbb{R}$, $f_L \leq f_R$, $g_L, g_R : \mathbb{R} \to [0, \infty)$, $h_L, h_R : \mathbb{R} \to (0, \infty)$, are Borel-measurable functions.

Given $X : \Omega \to \mathbb{R}$ a real random variable associated with $(\Omega, \mathcal{A}, P)$, it is straightforward to show that the mapping $\gamma \circ X : \Omega \to \mathcal{F}_c(\mathbb{R})$, $\omega \mapsto \gamma(X(\omega))$ is a random fuzzy set. It is called the *$\gamma$-fuzzy representation of* $X$ [8]. One of the main statistical advantages of this fuzzification process is the possibility of managing real-valued distributions, generally complicated in the classical framework, through powerful statistical techniques for random fuzzy variables which are available in the current literature on the fuzzy framework.

Several statistical problems for $X$ have been already solved by means of this technique [1, 3, 8, 9]. Different fuzzy operators $\gamma$ are considered, depending on the relevant information from $X$ which it is desired to characterize. There exists the possibility of characterizing the whole distribution of $X$ through the expected value of certain fuzzy representations. The fuzzy operator $\gamma^{\xi}$ is defined as

$$\gamma^{\xi}(x) = \mathbf{1}_{\{x\}} + \text{sig}(x - x_0)\gamma_f\left(\left|\frac{x - x_0}{a}\right|\right),$$

where $\xi \in \Theta = \{(x_0, a, f) | x_0 \in \mathbb{R}, a \in \mathbb{R}^+, \ f : [0, +\infty) \to [0, 1]$ injective and continuous$\}$, sig$(z)$ denotes the sign of $z \in \mathbb{R}$ and $\gamma_f : [0, +\infty) \to \mathcal{F}_c(\mathbb{R})$ is an auxiliary (fuzzy-valued) functional defined by

$$\left(\gamma_f(x)\right)_{\alpha} = \begin{cases} [0, B(x) - C(x)\alpha] & \text{if } 0 \le \alpha \le f(x) \\ [0, A(x)(1 - \alpha)] & \text{if } f(x) < \alpha \le 1 \end{cases}$$

for all $\alpha \in [0, 1]$ and $x \ge 0$, where

$$A(x) = \frac{x^2}{1 - f(x)}, \ B(x) = \frac{x^2}{f(x)} \ \text{ and } \ C(x) = \frac{x^2(1 - f(x))}{f(x)^2}.$$

The triple parameter $\xi = (E(X), 1, (0.6^x + 0.001)/1.001)$ provides a good exploratory analysis of $X$, as well as the characterization of its distribution, since two real random variables $X$ and $Y$ are identically distributed if, and only if, $\widetilde{E}(\gamma^{\xi} \circ X) = \widetilde{E}(\gamma^{\xi} \circ Y)$ (see [3]).

## 3 Testing the Equality of Real-Valued Distributions

Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $X_1, X_2, \ldots, X_k : \Omega \to \mathbb{R}$ be $k$ real-valued random variables. The aim is to test whether the distributions of the $k$ variables behave significantly different each other or not. Thus, the hypothesis test to be solved is:

$$\begin{cases} H_0 : X_1 \overset{d}{\sim} X_2 \overset{d}{\sim} \cdots \overset{d}{\sim} X_k \\ H_1 : \exists\, i, j \in \{1, \ldots, k\} \; s.t. \; X_i \overset{d}{\nsim} X_j \end{cases} \tag{1}$$

Following previous results on the fuzzy representation of the variables, it is immediate to note that the hypothesis test (1) can be equivalently written in terms of the expected values of the corresponding $\gamma^{\xi}$-fuzzy representations of the variables, as follows:

$$\begin{cases} H_0 : E(\gamma^{\xi} \circ X_1) = E(\gamma^{\xi} \circ X_2) = \cdots = E(\gamma^{\xi} \circ X_k) \\ H_1 : \exists\, i, j \in \{1, \ldots, k\} \; s.t. \; E(\gamma^{\xi} \circ X_i) \neq E(\gamma^{\xi} \circ X_j) \end{cases} \tag{2}$$

Whenever the normality assumption for the distributions of the variables is not guaranteed, the classical test (1) is solved through non-parametric techniques; $k$-sample Kolmogorov-Smirnov test, Kruskal-Wallis method, are some of the well-known alternatives, among others. Nevertheless, the equality of expectations of fuzzy random variables is tested through parametric techniques, shown to be asymptotically consistent. Let us define $\mathcal{X}_i = \gamma^\xi \circ X_i$, the $\gamma^\xi$-fuzzy representation of the real random variable $X_i$, respectively for $i = 1, \dots, k$. Given $\{X_{ij}\}_{j=1}^{n_i}$ a simple random sample from the real random variable $X_i$, for each $i = 1, \dots, k$, it is immediate to see that $\{\mathcal{X}_{ij} = \gamma^\xi \circ X_{ij}\}_{j=1}^{n_i}$ is a simple random sample from the random fuzzy variables $\mathcal{X}_i$, $i = 1, \dots, k$.

By following ideas from [9], the test statistic to be considered to solve (2) from the information provided by the random sample $\{\mathcal{X}_{ij}\}_{j=1}^{n_i}$ is defined as follows:

$$T_n = \sum_{i=1}^{k} n_i \left( D_\theta^\varphi(\overline{\mathcal{X}_{i\cdot}}, \overline{\mathcal{X}_{\cdot\cdot}}) \right)^2, \tag{3}$$

where $\overline{\mathcal{X}_{i\cdot}} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{X}_{ij}$ for each $i = 1, \dots, k$, $\overline{\mathcal{X}_{\cdot\cdot}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathcal{X}_{ij}$, and $n = n_1 + \cdots + n_k$ is the overall sample size. The consistency of the testing procedure based on the test statistic (3) is supported by the following asymptotic result.

**Theorem 1** (see [9]) *If $n_i \to \infty$, $n_i/n \to p_i > 0$, as $n \to \infty$, and $\mathcal{X}_i$ is non-degenerated for some $i \in \{1, \dots, k\}$, then, if $H_0$ is true,*

$$T_n \xrightarrow{n} \sum_{i=1}^{k} \left( ||Z_i - \sum_{l=1}^{k} \alpha_{li} Z_l||_\theta^\varphi \right)^2, \tag{4}$$

*where $Z_1, \dots, Z_k$ are independent centered Gaussian processes in $\mathcal{L}^2(\{-1, 1\} \times [0, 1])$ whose covariances are equal to $cov(s_{\mathcal{X}_i})$, respectively, and $\alpha_{li} = \sqrt{p_l/p_i} \sum_{r=1}^{k} (p_r/p_i)$, $i = 1, \dots, k$.*

**Proposition 1** *To test $H_0 : E(\gamma^\xi \circ X_1) = E(\gamma^\xi \circ X_2) = \cdots = E(\gamma^\xi \circ X_k)$ at the nominal significance level $\rho \in [0, 1]$, $H_0$ should be rejected whenever $T_n > z_\rho$, where $z_\rho$ is the $100(1 - \rho)$-quantile of the distribution of the limit expression in (4).*

The limit distribution in (4) depends on the populational covariances $cov(s_{\mathcal{X}_i})$, $i = 1, \dots, k$, which are usually unknown in practice. In such situations, that distribution can be approximated by Monte Carlo simulations.

Alternatively to the asymptotic approach, a bootstrap testing procedure to solve (2) is proposed, which is always applicable in practice. Let $\{\mathcal{X}_{ij}^*\}_{j=1}^{n_i}$ be a bootstrap sample of $\{\mathcal{X}_{ij}\}_{j=1}^{n_i}$, for each $i = 1, \dots, k$, i.e. $\{\mathcal{X}_{ij}^*\}_{j=1}^{n_i}$ being randomly chosen and with replacement from $\{\mathcal{X}_{ij}\}_{j=1}^{n_i}$. The bootstrap statistic is defined as follows:

$$T_n^* = \sum_{i=1}^{k} n_i \left( D_\theta^\varphi(\overline{\mathcal{X}_{i\cdot}^*} + \overline{\mathcal{X}_{\cdot\cdot}}, \overline{\mathcal{X}_{i\cdot}} + \overline{\mathcal{X}_{\cdot\cdot}^*}) \right)^2. \tag{5}$$

By applying the Bootstrap Central Limit Theorem, it can be shown that $T_n^*$ converges in law to the same Gaussian process than $T_n$ in (3) when $H_0$ is true (see [9]). Consequently, the bootstrap distribution of $T_n^*$ approximates the one of $T_n$ under $H_0$, and the following test resolution holds.

**Proposition 2** *To test* $H_0 : E(\gamma^\xi \circ X_1) = E(\gamma^\xi \circ X_2) = \cdots = E(\gamma^\xi \circ X_k)$ *at the significance level* $\rho \in [0, 1]$*,* $H_0$ *should be rejected whenever* $T_n^* > z_\rho^*$*, where* $z_\rho^*$ *is the* $100(1 - \rho)$*-quantile of the bootstrap distribution of* $T_n^*$*.*

The bootstrap statistic $T_n^*$ is defined coherently in terms of the arithmetic between fuzzy values and distances between them. The corresponding quantile to solve the test can be easily approximated by re-sampling.

## *3.1 Simulation Studies*

In order to illustrate the empirical behaviour of the proposed testing procedures, some simulations are shown. Let $\delta > 0$, $Z_i \hookrightarrow \mathcal{N}(0, 1)$, $i = 1, 2, 3$. We define $X_1 = Z_1$, $X_2 = Z_2$ and $X_3 = \delta Z_3$. It is immediate to check that $H_0 : X_1 \overset{d}{\sim} X_2 \overset{d}{\sim} X_3$ holds when $\delta = 1$. $E(X_1) = E(X_2) = E(X_3)$ for all $\delta > 0$. However, $Var(X_3)$ increases and it differs more and more from $Var(X_1) = Var(X_2)$ as $\delta$ increases.

A number of 10,000 random samples from $\{X_i\}_{i=1}^3$ are generated, for different sample sizes $n_1, n_2, n_3$, respectively. For each case, the corresponding samples for the fuzzy-representations $\gamma^\xi \circ X_i$ are constructed, with $\xi = (E(X), 1, (0.6^x + 0.001)/1.001)$, and the bootstrap test is run for $B = 1000$ bootstrap replications. The percentage of rejections of the null hypothesis (2) (and so of (1)) on the 10,000 iterations of the test is computed. The results are compared with the classical nonparametric Kruskal-Wallis (KW) method to test the equality of real-valued distributions. Table 1 contains the results for different values of $\delta$. Some comments can be done. First, it is immediate to see that the proposed bootstrap test approximates the nominal significance level when $H_0$ is true ($\delta = 1$) as the sample size increases, which agrees with the theoretical correctness of the method. Under $H_0$, the classical KW test approximates slightly better than the bootstrap test. However, the fuzzy-based bootstrap test is always consistent, which is not the case of the classical one.

**Table 1** Percentage of rejections (bootstrap fuzzy test/classical KW test)

| $(n_1, n_2, n_3)$ | $\boldsymbol{\delta = 1}$ | $\delta = 2$ | $\delta = 4$ | $\delta = 6$ |
|---|---|---|---|---|
| (30, 30, 30) | **3.80/4.69** | 34.16/5.70 | 34.46/6.81 | 35.86/7.45 |
| (30, 50, 100) | **4.22/5.09** | 90.36/2.46 | 71.62/2.52 | 83.32/2.62 |
| (100, 100, 100) | **4.62/5.13** | 90.52/5.59 | 72.22/7.20 | 88.45/7.68 |
| (100, 150, 200) | **4.68/5.04** | 97.78/4.13 | 87.64/4.92 | 99.48/5.15 |
| (500, 500, 500) | **4.80/5.11** | 99.98/5.88 | 100/6.96 | 100/7.89 |

Kruskal-Wallis method does not identify the movement of the theoretical situation far from $H_0$ (when $\delta$ increases), whereas the bootstrap method does it effectively even for small and moderate samples. Despite the fact that the KW test is used as a non-parametric test to check the equality of distributions, it is, in fact, a test for the comparison of medians. This could be a reason to the inconsistency of the KW test when $\delta > 1$.

### 3.2 Practical Applications

Once both the theoretical and empirical correctness of the proposed fuzzy-based bootstrap testing procedure is shown, the technique is ready to be applied in practice. Let us consider the sample dataset *Energy Efficiency* from the UCI Repository (see [15]). It contains information about the heating load of houses as well as different features of the houses such as orientation, roof area, wall area, glazing area, etc. The aim is to test whether the heating load of the houses is significantly different depending on one of those features. For instance, if we consider $X_i$ = heating load in a house with orientation $i, i = 1, 2, 3, 4$, the hypothesis $H_0 : X_1 \overset{d}{\sim} X_2 \overset{d}{\sim} X_3 \overset{d}{\sim} X_4$ is not rejected at the usual significance levels, since the obtained p-values with both the classical KW test and the fuzzy bootstrap test are 0.9941 and 0.8741, respectively. Nevertheless, for $Y_j$ = heating load in a house with glazing area $j, j = 1, 2, 3$, the hypothesis $H_0 : Y_1 \overset{d}{\sim} Y_2 \overset{d}{\sim} Y_3$ is rejected with p-values $4.31 \times 10^{-13}$ and 0, respectively.

## 4 Conclusions

The statistical analysis of fuzzy-valued random sets is widely recognized as a powerful technique to develop descriptive and inferential studies in experimental scenarios executed with certain degree of imprecision. Additionally, fuzzy-based techniques can be also applied to solve statistical problems for real-valued random variables through the fuzzy representation of the variables. In this work, the problem of testing the equality of two or more real-valued distributions is addressed. Simulations and applications show that the proposed techniques are a good alternative to classical methods, with good and powerful statistical features.

The effect of using different fuzzy-operators to define the fuzzy representation of the variables in the results of the test could be further investigated. Besides, the extension to other classical statistical methods, as discriminant or regression analysis, is still to be developed.

**Acknowledgments** The research in this paper is partially supported by the Spanish National Grant MTM2013-44212-P, and the Regional Grant FC-15-GRUPIN-14-005. Their financial support is gratefully acknowledged.

## References

1. Blanco-Fernández A, Ramos-Guajardo AB, Colubi A (2013) Fuzzy representations of real-valued random variables: applications to exploratory and inferential studies. Metron 71:245–259
2. Blanco-Fernández A, Casals RM, Colubi A, Corral N, García-Bárzana M, Gil MA, González-Rodríguez G, López MT, Lubiano MA, Montenegro M, Ramos-Guajardo AB, de la Rosa de Sáa S, Sinova B, (2014) A distance-based statistical analysis of fuzzy number-valued data. Int J Approx Reason 55:1487–1501
3. Colubi A (2009) Statistical inference about the means of fuzzy random variables: applications to the analysis of fuzzy- and real-valued data. Fuzzy Set Syst 160:344–356
4. Couso I, Dubois D (2009) On the variability of the concept of variance for fuzzy random variables. IEEE Trans Fuzzy Sys 17(5):1070–1080
5. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: ontic vs. epistemic views. Int J Approx Reason 55:1502–1518
6. Diamond P, Kloeden P (1994) Metric spaces of fuzzy sets. World Scientific, Singapore
7. Gil MA, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. Fuzzy Set Syst 157(19):2546–2557
8. González-Rodríguez G, Colubi A, Gil MA (2006) A fuzzy representation of random variables: an operational tool in exploratory analysis and hypothesis testing. Comput Stat Data Anal 51:163–176
9. González-Rodríguez G, Colubi A, Gil MA, Lubiano MA (2012) A new way of quantifying the symmetry of a random variable: estimation and hypothesis testing. J Stat Plann Inf 142:3061–3072
10. Huellermeier E (2014) Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int J Approx Reason 55:1519–1534
11. Körner R, Nather W (2002) On the variance of random fuzzy variables. In: Bertoluzza C, Gil MA, Ralescu DA (eds) Analysis and management of fuzzy data. Physica-Verlag, Heidelberg, pp 22–39
12. Kwakernaak H (1978) Fuzzy random variables—I. Definitions and theorems. Inf Sci 15:1–29
13. Puri ML, Ralescu DA (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
14. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inf Sci 179:3964–3972
15. Tsanas A, Xifara A (2012) Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Buildings 49:560–567
16. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8:199–249

# Agglomerative Fuzzy Clustering

**Christian Borgelt and Rudolf Kruse**

**Abstract** The term *fuzzy clustering* usually refers to prototype-based methods that optimize an objective function in order to find a (fuzzy) partition of a given data set and are inspired by the classical $c$-means clustering algorithm. Possible transfers of other classical approaches, particularly hierarchical agglomerative clustering, received much less attention as starting points for developing fuzzy clustering methods. In this chapter we strive to improve this situation by presenting a (hierarchical) agglomerative fuzzy clustering algorithm. We report experimental results on two well-known data sets on which we compare our method to classical hierarchical agglomerative clustering.

## 1 Introduction

The objective of *clustering* or *cluster analysis* is to divide a data set into groups (so-called *clusters*) in such a way that data points in the same cluster are as similar as possible and data points from different clusters are as dissimilar as possible (see, e.g., [5, 10]), where the notion of similarity is often formalized by defining a distance measure for the data points. Even in classical clustering the resulting grouping need not be a partition (that is, in some approaches not all data points need to be assigned to a group and the formed groups may overlap), but only if points are assigned to different groups with different degrees of membership, one arrives at *fuzzy clustering* [2, 3, 8, 14].

However, the term *fuzzy clustering* usually refers to a fairly limited set of methods, which are prototype-based and optimize some objective function to find a good (fuzzy) partition of the given data. Although classical clustering comprises many

C. Borgelt (✉) · R. Kruse
School of Computer Science, Otto-von-Guericke-University Magdeburg,
Universitätsplatz 2, 39106 Magdeburg, Germany
e-mail: christian@borgelt.net

R. Kruse
e-mail: rudolf.kruse@ovgu.de

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_9

more methods than the well-known $c$-means algorithm (by which most fuzzy clustering approaches are inspired), these other methods are only rarely “fuzzified”. This is particularly true for hierarchical agglomerative clustering (HAC) [16], of which only few fuzzy versions have been proposed.

Exceptions include [1, 7, 11]. Ghasemigol et al. [7] describes HAC for trapezoidal fuzzy sets with either single or complete linkage, but is restricted to one dimension due to its special distance function. Konkol [11] proposes an HAC algorithm for crisp data based on fuzzy distances, which are effectively distances weighted by a function of membership degrees. It mixes single and complete linkage. Bank and Schwenker [1] merges clusters in the spirit of HAC, but keeps the original clusters for possible additional mergers, so that a hierarchy in the form of a directed acyclic graph results (while standard HAC produces a tree). Also noteworthy is [15], which suggest a mixed approach, re-partitioning the result of fuzzy $c$-means clustering and linking the partitions of two consecutive steps.

Related approaches include [6, 12] as well as its extension [9]. The first uses a competitive agglomeration scheme and an extended objective function for fuzzy $c$-means in order to reduce an overly large initial number of clusters to an “optimal” number. The latter two change the term in the objective function that penalizes many clusters from a quadratic expression to an entropy expression. Although fairly different from hierarchical agglomerative clustering approaches, they share the property that clusters are merged to find a good final partition, but they do not necessarily produce a hierarchy.

Our approach is closest in spirit to [11], as it also relies on the standard scheme of hierarchical agglomerative clustering, although we treat the original data points as clusters already, while [11] keeps data points and clusters clearly separate. Furthermore, [11] focuses on single and complete linkage while we use a centroid scheme. Our approach also bears some relationship to [15] concerning the distances of fuzzy sets, which [15] divides into three categories: (1) comparing membership values, (2) considering spatial characteristics, and (3) characteristic indices. While [15] relies on (2), we employ (1).

The remainder of this paper is structured as follows: in Sects. 2 and 3 we briefly review standard fuzzy clustering and hierarchical agglomerative clustering, indicating which elements we use in our approach. In Sect. 4 we present our method and in Sect. 5 we report experimental results. Finally, we draw conclusions from our discussion in Sect. 6.

## 2 Fuzzy Clustering

The input to our clustering algorithm is a data set $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with $n$ data points, each of which is an $m$-dimensional real-valued vector, that is, $\forall j; 1 \le j \le n$ : $\mathbf{x}_j = (x_{j1}, \dots, x_{jm}) \in \mathbb{R}^m$. Although HAC usually requires only a distance or similarity matrix as input, we assume metric data, since a centroid scheme requires the possibility to compute new center vectors.

In standard fuzzy clustering one tries to minimize the objective function

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{w} \, d_{ij}^{2},$$

where $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_c\}$ is the set of *prototypes* (often merely cluster centers), the $c \times n$ matrix $\mathbf{U} = (u_{ij})_{1 \leq i \leq c; 1 \leq j \leq n}$ is the *partition matrix* containing the degrees of membership with which the data points belong to the clusters, the $d_{ij}$ are the distances between cluster $\mathbf{c}_i$ and data point $\mathbf{x}_j$, and $w$, $w > 1$, is the *fuzzifier* (usually $w = 2$), which controls the "softness" of the cluster boundaries (the larger $w$, the softer the cluster boundaries). In order to avoid the trivial solution of setting all membership degrees to zero, the constraints $\forall j; 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$ and $\forall i; 1 \leq i \leq c : \sum_{j=1}^{n} u_{ij} > 0$ are introduced.

A fuzzy clustering algorithm optimizes the above function, starting from a random initialization of either the cluster prototypes or the partition matrix, in an alternating fashion: (1) optimize membership degrees for fixed prototypes and (2) optimize prototypes for fixed membership degrees. From this scheme we take the computation of membership degrees for $w = 2$, namely

$$u_{ij} = \frac{d_{ij}^{-2}}{\sum_{k=1}^{c} d_{kj}^{-2}}.$$

We compute membership degrees for fuzzy clusters in an only slightly modified fashion, which are then compared to decide which clusters to merge.

## 3 Hierarchical Agglomerative Clustering

As its name already indicates, *hierarchical agglomerative clustering* produces a hierarchy of clusters in an agglomerative fashion, that is, by merging clusters (in contrast to divise approaches, which split clusters). It starts by letting each data point form its own cluster and then iteratively merges those two clusters that are most similar (or closest to each other).

While the similarity (or distance) of the data points is an input to the procedure, how the distances of (non-singleton) clusters are to be measured is a matter of choice. Common options include (1) *single linkage* (cluster distances are minimum distances of contained data points), (2) *complete linkage* (maximum distances of contained data points), and (3) *centroid* (distances of cluster centroids). Note that the centroid method requires that one can somehow compute a cluster center (or at least an analog), while single and complete linkage only require the initial similarity or distance matrix of the data points. Because of this we assume metric data as input.

In the single and complete linkage methods, clusters are merged by simply pooling the contained data points. In the centroid method, clusters are merged by computing a new cluster centroid as the weighted mean of the centroids of the clusters to be merged, where the weights are provided by the relative number of data points in the clusters to be merged.

## 4 Agglomerative Fuzzy Clustering

Our algorithm builds on the idea to see the given set of data points as the initial cluster centers (as in standard HAC) and to compute membership degrees of all data points to these cluster centers. However, for this the membership computation reviewed in Sect. 2 is not quite appropriate, since it leads to each data point being assigned to itself and to itself only (only one membership degree is 1, all others are 0). As a consequence, there would be no similarity between any two clusters (at least in the initial partition) and thus no proper way to choose a cluster merger. In order to circumvent this problem, we draw on the concept of a "raw" membership degree, which is computed from a distance via a radial function, where "raw" means that its value is not normalized to sum 1 over the clusters [4]. Possible choices for such a radial function (with parameters $\alpha$ and $\sigma^2$, respectively) are

$$f_{\text{Cauchy}}(r; \alpha) = \frac{1}{r^2 + \alpha} \qquad \text{and} \qquad f_{\text{Gauss}}(r; \sigma^2) = e^{-\frac{r^2}{2\sigma^2}},$$

where $r$ is the distance to a cluster center. Using these functions (with $\alpha > 0$) prevents singularities at the cluster centers that occur with the simple inverted squared distance (that is, for $\alpha = 0$) and thus allows us to compute suitable membership degrees even for the initial set of clusters, namely as

$$u_{ij}^{(\alpha)} = \frac{f_{\text{Cauchy}}(d_{ij}; \alpha)}{\sum_{k=1}^{c} f_{\text{Cauchy}}(d_{kj}; \alpha)} \qquad \text{or} \qquad u_{ij}^{(\sigma^2)} = \frac{f_{\text{Gauss}}(d_{ij}; \sigma^2)}{\sum_{k=1}^{c} f_{\text{Gauss}}(d_{kj}; \sigma^2)}.$$

Based on these membership degrees two clusters $\mathbf{c}_i$ and $\mathbf{c}_k$ can now be compared by aggregating (here: simply summing) point-wise comparisons:

$$\delta_{ik} = \sum_{j=1}^{n} g(u_{ij}, u_{kj}),$$

where $g$ is an appropriately chosen difference function. Here we consider

$$g_{\text{abs}}(x, y) = |x - y|, \quad g_{\text{sqr}}(x, y) = (x - y)^2 \quad \text{and} \quad g_{\text{wgt}}(x, y) = (x - y)(x + y).$$

The first function, $g_{\text{abs}}$, may appear the most natural choice, while $g_{\text{sqr}}$ generally weights large differences more strongly and $g_{\text{wgt}}$ emphasizes large differences of large membership degrees and thus focuses on close neighbors.

A fuzzy HAC algorithm can now be derived in a standard fashion: compute the initial membership degrees by using each data point as a cluster center. Compute the cluster dissimilarities $\delta_{ik}$ for this initial set of clusters. Merge the two clusters $\mathbf{c}_i$ and $\mathbf{c}_k$, for which $\delta_{ik}$ is smallest, according to

$$\mathbf{c}_* = \frac{1}{\sum_{j=1}^{n}(u_{ij} + u_{kj})}\left(\mathbf{c}_i \sum_{j=1}^{n} u_{ij} + \mathbf{c}_k \sum_{j=1}^{n} u_{kj}\right).$$

That is, the sum of membership degrees for each cluster is used as the relative weight of the cluster for the merging and thus (quite naturally) replaces the number of data points in the classical HAC scheme. The merged clusters $\mathbf{c}_i$ and $\mathbf{c}_k$ are removed and replaced by the result $\mathbf{c}_*$ of the merger.

For the next step membership degrees and cluster dissimilarities are re-computed and again the two least dissimilar clusters are merged. This process is repeated until only one cluster remains. From the resulting hierarchy a suitable partition may then be chosen to obtain a final result (if so desired), which may be further optimized by applying standard fuzzy $c$-means clustering.

Note that this agglomerative fuzzy clustering scheme is computationally considerably more expensive than standard HAC, since all membership degrees and cluster dissimilarities need to be re-computed in each step.

## 5 Experimental Results

We implemented our agglomerative fuzzy clustering method prototypically in Python, allowing for the two radial functions (Cauchy and Gauss, with parameters $\alpha$ and $\sigma^2$) to compute membership degrees and the three cluster dissimilarity measures ($g_{\text{abs}}$, $g_{\text{sqr}}$ and $g_{\text{wgt}}$) to decide which clusters to merge. We applied this implementation in a simple first test of functionality to two well-known data sets from the UCI machine learning repository [13], namely the Iris data and the Wine data. For the clustering runs we used attributes petal_length and petal_width for the Iris data and attributes 7, 10 and 13 for the Wine data, since these are the most informative attributes w.r.t. the class structure of these data sets. This restriction of the attributes also allows us to produce (low-dimensional) diagrams with which the cluster hierarchies can be easily compared. The latter is important, since it is difficult to find an undisputed way of evaluating clustering results. Visual representations in diagrams at least allow to compare the results subjectively and provide some insight about the properties of the different variants.

As a comparison we applied standard hierarchical agglomerative clustering (HAC) with the centroid method for linking clusters. As it also produces a hierarchy of clusters (cluster tree) the results can be displayed in the same manner and thus are easy to compare. For both standard HAC and agglomerative fuzzy clustering we $z$-normalized the data (that is, we normalized each attribute to mean 0 and standard deviation 1) in order to avoid effects resulting from different scales (which is particularly important for attribute 13 of the Wine data set, which spans a much larger range than all other attributes and thus would dominate the clustering without normalization).

A selection of results we obtained are shown in Figs. 1 and 2 for the Iris data and in Figs. 3 and 4 for the Wine data. Since in our approach cluster dissimilarity basically depends on all data points, the distribution of the data points in the data space has

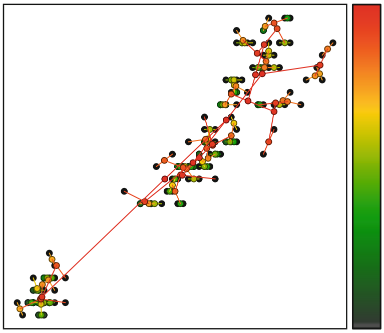
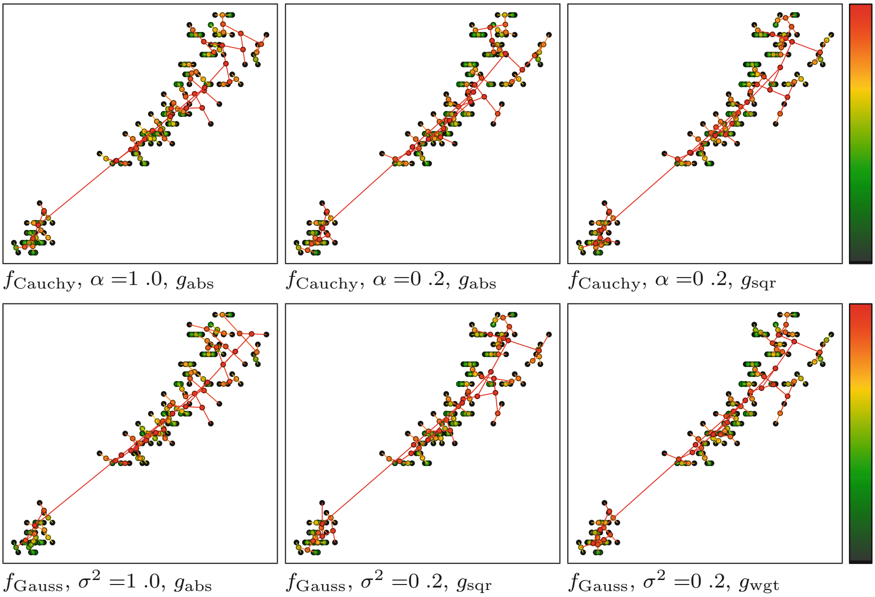


**Fig. 1** Result of standard hierarchical agglomerative clustering (i.e. crisp partitions) with the centroid method on the well-known Iris data, attributes petal_length (*horizontal*) and petal_width (*vertical*). The colors encode the step in which clusters are merged (from *bottom* to *top* on the *color bar* shown on the *right*); the data points are shown in *gray*



$f_{\text{Cauchy}}$, $\alpha = 1.0$, $g_{\text{abs}}$ $f_{\text{Cauchy}}$, $\alpha = 0.2$, $g_{\text{abs}}$ $f_{\text{Cauchy}}$, $\alpha = 0.2$, $g_{\text{sqr}}$

$f_{\text{Gauss}}$, $\sigma^2 = 1.0$, $g_{\text{abs}}$ $f_{\text{Gauss}}$, $\sigma^2 = 0.2$, $g_{\text{sqr}}$ $f_{\text{Gauss}}$, $\sigma^2 = 0.2$, $g_{\text{wgt}}$

**Fig. 2** Results of different versions of agglomerative fuzzy clustering on the Iris data, attributes petal_length (*horizontal*) and petal_width (*vertical*)

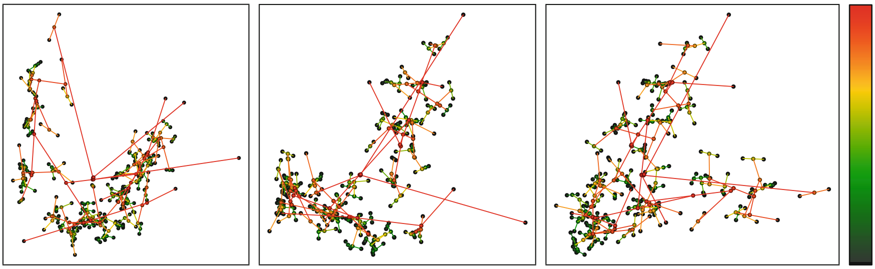
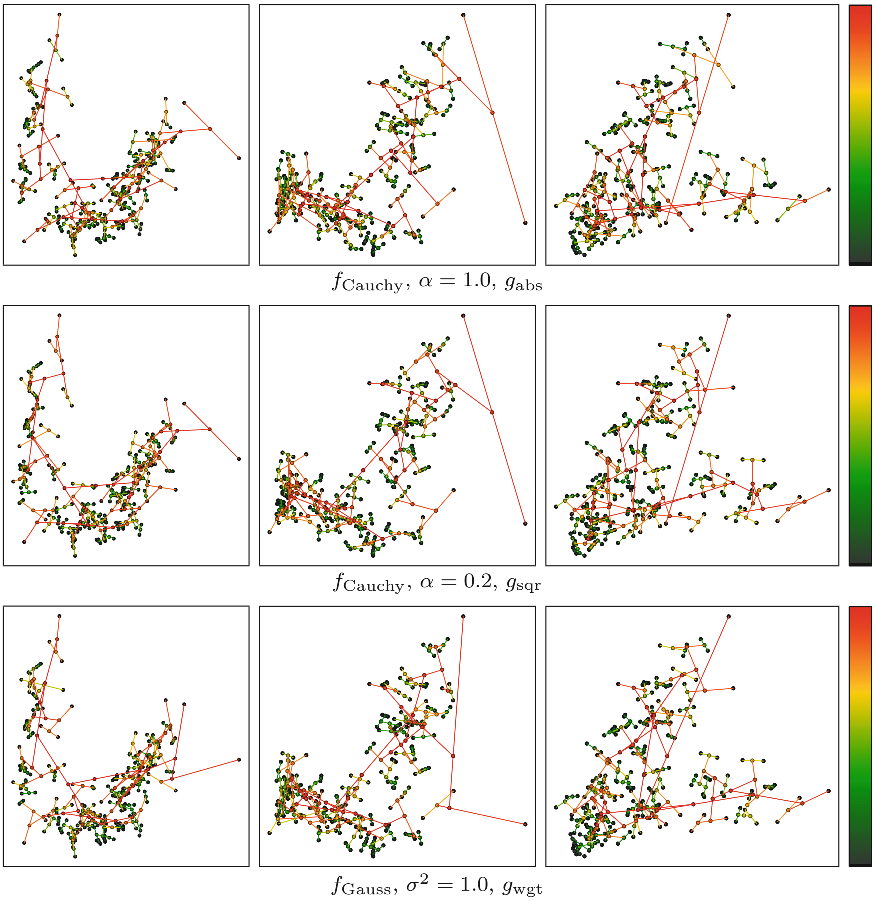


**Fig. 3** Result of standard hierarchical agglomerative clustering (i.e. crisp partitions) with the centroid method on the Wine data, attributes 7 and 10 (*left*), 7 and 13 (*middle*) and 10 and 13 (*right*). The *colors* encode the step in which clusters are merged (from *bottom* to *top* on the *color bar* shown on the *right*); the data points are shown in *gray*



$f_{\text{Cauchy}},\ \alpha = 1.0,\ g_{\text{abs}}$

$f_{\text{Cauchy}},\ \alpha = 0.2,\ g_{\text{sqr}}$

$f_{\text{Gauss}},\ \sigma^2 = 1.0,\ g_{\text{wgt}}$

**Fig. 4** Results of different versions of agglomerative fuzzy clustering on the Wine data, projections to attributes 7 and 10 (*left*), 7 and 13 (*middle*) and 10 and 13 (*right*)

a stronger influence on the mergers to be carried out. For example, for the Iris data, which is mainly located along a diagonal of the data space, mergers with our algorithm tend to be carried out more often in a direction perpendicular to this diagonal. How strong this effect is depends on the parameters: a smaller $\alpha$ or $\sigma^2$ reduces this effect. For the wine data set, which has a more complex data distribution, we believe that we can claim that the resulting cluster trees better respects the distribution of the data points than standard HAC does.

## 6 Conclusions

We described a (hierarchical) agglomerative fuzzy clustering algorithm (fuzzy HAC) that is based on a cluster dissimilarity measure derived from aggregated point-wise membership differences. Although it is computationally more expensive than classical (crisp) HAC, a subjective evaluation of its results seems to indicate that it may be able to produce cluster hierarchies that better fit the spatial distribution of the data points than the hierarchy obtained with classical HAC. Future work includes a more thorough investigation of the effects of its parameters ($\alpha$ and $\sigma^2$ and the choice of the dissimilarity function, as well as the fuzzifier, which we neglected in this paper). Furthermore, an intermediate (partial) optimization of the cluster centers with fuzzy $c$-means is worth to be examined and may make it possible to return to simple inverted squared distances to compute the membership degrees.

## References

1. Bank M, Schwenker F (2012) Fuzzification of agglomerative hierarchical crisp clustering algorithms. In: Proceedings 34th annual conference on Gesellschaft für Klassifikation (GfKl, (2010) Karlsruhe, Germany). Springer, Heidelberg/Berlin, Germany, pp 3–11
2. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
3. Bezdek JC, Keller J, Krishnapuram R, Pal N (1999) Fuzzy models and algorithms for pattern recognition and image processing. Kluwer, Dordrecht
4. Borgelt C (2005) Prototype-based classification and clustering. Otto-von-Guericke-University of Magdeburg, Germany, Habilitationsschrift
5. Everitt BS (1981) Cluster analysis. Heinemann, London
6. Frigui H, Krishnapuram R (1997) Clustering by competitive agglomeration. Pattern Recogn 30(7):1109–1119. Elsevier, Amsterdam, Netherlands
7. Ghasemigol M, Yazdi HS, Monsefi R (2010) A new hierarchical clustering algorithm on fuzzy data (FHCA). Int J Comput Electr Eng 2(1):134–140. International Academy Publishing (IAP), San Bernadino, CA, USA
8. Höppner F, Klawonn F, Kruse R, Runkler T (1999) Fuzzy cluster analysis. Wiley, Chichester
9. Hsu M-J, Hsu P-Y, Dashnyam B (2011) Applying agglomerative fuzzy K-means to reduce the cost of telephone marketing. In: Proceedings of international symposium on integrated uncertainty in knowledge modelling and decision making (IUKM 2011, Hangzhou, China), pp 197–208

10. Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York
11. Konkol M (2015) Fuzzy agglomerative clustering. In: Proceedings of 14th international conference on artificial intelligence and soft computing (ICAISC, (2015) Zakopane, Poland). Springer, Heidelberg/Berlin, Germany, pp 207–217
12. Li M, Ng M, Cheung Y-M, Huang J (2008) Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters. IEEE Trans Knowl Data Eng 20(11):1519–1534. IEEE Press, Piscataway, NJ, USA
13. Lichman M (2013) UCI machine learning repository. University of California, Irvine, CA, USA. http://archive.ics.uci.edu/ml
14. Ruspini EH (1969) A new approach to clustering. Inf Control 15(1):22–32. Academic Press, San Diego, CA, USA
15. Torra V (2005) Fuzzy c-means for fuzzy hierarchical clustering. In: Proceedings of IEEE international conference on fuzzy systems (FUZZ-IEEE, (2005) Reno, Nevada). IEEE Press, Piscataway, NJ, USA, pp 646–651
16. Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244

# Bayesian Inference for a Finite Population Total Using Linked Data

**Dario Briscolini, Brunero Liseo and Andrea Tancredi**

**Abstract** We consider the problem of estimating the total (or the mean) of a continuous variable in a finite population setting, using the auxiliary information provided by a covariate which is available in a different file. However the matching steps between the two files is uncertain due to a lack of identification code for the single unit. We propose a fully Bayesian approach which merges the record linkage step with the subsequent estimation procedure.

## 1 Introduction

Statistical Institutes and other private and public agencies often need to integrate statistical knowledge extracting information from different sources. This operation may be important for two different and complementary reasons:

(i) per sé, i.e. to obtain a larger reference data set or frame, suitable to perform more accurate statistical analyses;
(ii) to calibrate statistical models via the additional information which could not be extracted from either one of the single data sets.

When the merging step can be accomplished without errors there are no practical consequences. In practice, however, identification keys are rarely available and linkage among different records is usually performed under uncertainty. This issue has caused a very active line of research among the statistical and the machine learning communities, named "record linkage", or "data matching". In these approaches one

D. Briscolini · B. Liseo · A. Tancredi (✉)
MEMOTEF, Sapienza Università di Roma, Rome, Italy
e-mail: andrea.tancredi@uniroma1.it

D. Briscolini
e-mail: dario.briscolini@uniroma1.it

B. Liseo
e-mail: brunero.liseo@uniroma1.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_10

should consider the possibility to make wrong matching decisions, especially when the result of the linking operation, namely the merged data set, must be used for further statistical analyses.

In this short note there is no room to extensively describe what record linkage is and how it is implemented. So we limit ourselves to a sketch. Suppose we have two data sets, say $A$ and $B$, whose records relate to statistical units (e.g. individuals, firms, etc.) of partially overlapping samples (or populations), say $S_A$ and $S_B$. Records in each data set consist of several fields, or variables, either quantitative or categorical, which may be observed together with a potential amount of noise. For example, in a file of individuals, fields could be *surname, age, sex*, and so on.

The goal of a record linkage procedure is to detect all the pairs of units $(j, j')$, with $j \in S_A$ and $j' \in S_B$, such that $j$ and $j'$ actually refer to the same individual. If the main goal of the record linkage process is the former outlined above (case (i)), a new data set is created by merging together three different subsets of units: those which are in both data sets and those belonging to $S_A$ $(S_B)$ only. Suitable data analyses may be then performed on the enlarged data set. Since the linkage step is done with uncertainty, the performance and the reliability of the statistical analysis may be jeopardized by the presence of duplicate units and by a loss of power, mainly due to erroneous matching in the merging process.

The latter situation (case (ii)) is even more challenging: to fix the ideas, assume that the observed variables in $A$ are $(Y, X_1, X_2, \dots, X_r)$, and the observed variables in $B$ are $(Z, X_1, X_2, \dots, X_r)$. One might be interested in performing a linear regression analysis (see [6]) (or any other more complex association model) between $Y$ and $Z$, restricted to those pairs of records which are considered matches after the record linkage based on $(X_1, \dots, X_r)$. The difficulties in such a simple problem are discussed in [2–4]. In a regression example discussed in [6], it might be seen that the presence of false matches reduces the level of association between $Y$ and $Z$ and, as a consequence, they introduces a bias effect towards zero when estimating the slope of the regression line. Similar biases may appear in any statistical procedure and, in most of the cases, the bias takes a specific direction. In this paper we propose a Bayesian encompassing approach, where the posterior distribution of the quantity of interest intrinsically takes into account the matching step uncertainty. In particular, we consider the problem of estimating the total of a continuous variable $Y$ in a finite population framework, when data—possibly linked with error—are available on another continuous variable $Z$. The method could be easily extended to more than one response variable or covariate. In this set-up we consider the two-fold objective of (i) using the key variables $X_1, X_2, \dots, X_r$ to infer about the common units between sources $A$ and $B$ and, at the same time, (ii) adopting a model $\mathcal{M}$ to perform a statistical analysis based on $Y$ and $Z$ (or even including the common variables $X_i$'s), restricted to those records which have been recognized as matches. In order to pursue this goal, we propose a fully Bayesian analysis which is able—in a natural way—to

- improve the performance of the linkage step (through the use of the extra information contained in the $Y$'s and $Z$'s. This happens because pairs of records which do not adequately fit the model $\mathcal{M}$ will be automatically down-weighted in the matching estimation;

- account for matching uncertainty in the estimation procedure related to model $\mathcal{M}$ involving $Y$ and $Z$.
- improve the accuracy of the estimators of the parameters of model $\mathcal{M}$ in terms of bias.

In the next section we will briefly recall the Bayesian approach to record linkage of [5, 6]. In Sect. 3 we describe how to include the "estimation of total" step and we will compare the new proposal with the classical GREG technique. We illustrate our proposal with simulated data sets in the final section.

## 2 Record Linkage

Given two data sets, say $A$ and $B$, we observe, on $A$, records $(Y_i, X_{i,1}, \ldots, X_{i,r})$, $i = 1, 2, \ldots, n_A$; on $B$ we observe $(X_{j,1}, \ldots, X_{j,r}, Z_{j,1}, \ldots, Z_{j,k})$, $j = 1, \ldots, n_B$. Here variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k$ are potentially related to $\boldsymbol{Y}$ and they might be used in a statistical model. Variables $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_r)$ are called the key variables and they are used in order to identify common records between the two dataset. The key variables are usually categorical. We assume there are $G_l$ categories for $\boldsymbol{X}_l$. Extension to continuous variables is certainly possible (see [7] for a practical illustration). Let us denote $\boldsymbol{X}^A$ $(\boldsymbol{X}^B)$ the data matrix $n_A \times r$ $(n_B \times r)$. We introduce a latent matching matrix $C$, with $n_A$ rows and $n_B$ columns: each element of $C$ may be either 0 or 1; $C_{ij} = 1$ if the $i$th record of $\boldsymbol{X}^A$ and the $j$th record of $\boldsymbol{X}^B$ correspond to the same unit. The main goal of a record linkage analysis is the estimation of $C$. Related to this, it may be important to make inference on some statistical relationships between $\boldsymbol{Y}$ and the $\boldsymbol{Z}$'s, restricted to the matched pairs of records. Let $T$ be a $r \times g$ matrix where $g$ is the number of categories that each key variable can assume (in practice, each variable has a different $g$, so we can pick the highest value): each element of $T$ is the probability that a generic key variable assumes that generic category. We also assume that the key variables might be observed with error; for $l = 1, \ldots, r$ there is a probability $\gamma_l$ that the generic value of $\boldsymbol{X}_l$ is correctly observed. This can be formalized in a "hit-and miss" model [1]: let $(\tilde{\boldsymbol{X}}^A, \tilde{\boldsymbol{X}}^B)$ the true unobserved values of the key variables on the sample units.

$$p(\boldsymbol{X}^A, \boldsymbol{X}^B | \tilde{X}^A, \tilde{X}^B) = \prod_{dul} p(x_{dul} | \tilde{x}_{dul}, \gamma_l) = \prod_{dul} [\gamma_l I(x_{dul} = \tilde{x}_{dul}) + \frac{1 - \gamma_l}{G_l}], \tag{1}$$

where $d = \{A, B\}$, $u = \{(i = 1, 2, \ldots, n_A), (j = 1, 2, \ldots, n_B)\}$, $l = \{1, 2, \ldots, r\}$ and $x_{dul}$, $\tilde{x}_{dul}$ are the observed and the true value of the key variable $l$ of unit $u$ in the dataset $d$, respectively. The other part of the model is related to the true values of the key variables and it depends on the matrix $C$. In the following we will consider the parameters $T$ and $\gamma$ as fixed and known. The method can be easily extended to consider both the above parameters as unknown. We have the following structure:

$$p(\tilde{X}^A, \tilde{X}^B|C) = \prod_{i:C_{i,j}=0,\forall j} p(\tilde{x}_{Ai}) \prod_{j:C_{i,j}=0,\forall i} p(\tilde{x}_{Bj}) \prod_{i,j:C_{i,j}=1} p(\tilde{x}_{Ai}, \tilde{x}_{Bj})$$

where

$$p(\tilde{x}_{du}) = \prod_{\substack{s_1,s_2,\ldots,s_r: \\ T[s_1,s_2,\ldots,s_r]>0}} T[s_1, s_2, \ldots, s_r]^{I(\tilde{x}_{du}=(s_1,s_2,\ldots s_r))}$$

and

$$p(\tilde{x}_{Ai}, \tilde{x}_{Bj}) = \prod_{\substack{s_1,s_2,\ldots,s_r: \\ T[s_1,s_2,\ldots,s_r]>0}} T[s_1, s_2, \ldots, s_r]^{I(\tilde{x}_{Ai}=(s_1,s_2,\ldots s_r))}$$

if $\tilde{x}_{Ai} \neq \tilde{x}_{Bj}$ otherwise the last probability is 0. In the expressions above $s_1, s_2, \ldots, s_r$ are the generic categories assumed respectively by the key variables $1, 2, \ldots, r$ and $T[s_1, s_2, \ldots, s_r]$ is the probability extracted by the $T$ matrix and corresponding to the selected categories.

Assuming the independence among the key variables one has

$$T[s_1, s_2, \ldots, s_r] = \prod_{l=1}^{r} T_{l,s_l}$$

where $T_{l,s_l}$ is the element in position $(l, s_l)$ of the $T$ matrix. The prior on C is chosen to be uniform over the space of all possible matrices whose elements are either 0 or 1 and no more than one 1 is present in each row or column. To reproduce the posterior distribution of $C$ we used a Metropolis Hastings algorithm where, at each iteration, the *proposed* matrix is obtained from the current one by (1) deleting a match; (2) adding a match; (3) switching two existing matches.

## 3 Inference on Linked Data

From a frequentist point of view one of the most popular ways to estimate totals is the GREG estimator. Let us consider a variable $Y$ and suppose a sample of size $n$ from a finite population of size $N$ is available: the goal is to estimate $\sum_{i=1}^{N} y_i$. Suppose it is available, for all the population units the values of some covariate $z$. In this case one can construct an estimator which uses the auxiliary information provided by $z$. In detail, we assume that a $N \times k$ matrix $\boldsymbol{Z}$ is available, where $k$ is the number of covariates. Let $\boldsymbol{s} = \sum_{i=1}^{N} \boldsymbol{z_i}$, with $\boldsymbol{z_i} = (z_{i,1}, z_{i,2}, \ldots z_{i,k})'$ and let $\hat{\boldsymbol{s}} = \sum_{i=1}^{n} \boldsymbol{z_i}/\pi_i$, where the $\pi_i$'s are the inclusion probabilities. Let $\hat{y} = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i$; the GREG estimator for the total is defined as

$$\hat{Y}_{GR} = \hat{y} + (\boldsymbol{s} - \hat{\boldsymbol{s}})'\hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \frac{1}{\pi_i} \boldsymbol{z_i} \boldsymbol{z_i}'\right)^{-1} \sum_{i=1}^{n} \frac{1}{\pi_i} \boldsymbol{z_i} y_i$. When the matching step between $Y$'s and $z$ is not certain, one needs to produce a point estimate of $C$ first and then compute the GREG estimator, conditionally on the estimated linked pairs.

### 3.1 Linkage and Estimation of Total: Full Bayesian Approach

Consider the record linkage framework explained in Sect. 2 and assume that $\sum_{i=1}^{N} z_i$ is known and $Y_i | \boldsymbol{z_i} \sim N(\boldsymbol{z_i'}\boldsymbol{\beta}, \sigma^2)$, independently of each other. We also assume that $Z_i \sim p_z(\cdot)$. The choice of $p_z(\cdot)$ is not crucial and in this paper we will take it as a Gaussian distribution. Let $S^A$ ($S^B$) be the set of units contained in sample $A$ ($B$). One can see that

$$\sum_{i=1}^{N} y_i = \sum_{i \in S^A} y_i + \sum_{i \notin S^A} y_i = \sum_{i \in S^A} y_i + Y^*.$$

We need to produce a sample from the posterior distribution of $Y^*$, say $\pi(Y^* | \boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, Z_{S^B})$ where $\boldsymbol{X}^A$ ($\boldsymbol{X}^B$) is the matrix of key variables observed in sample $A$ ($B$) and $Y_{S^A}$ is the $n_A$-dimensional response vector related to sample $A$. Finally, $\boldsymbol{Z}_{S^B}$ is the $n_B \times k$ matrix of the potential covariates in $S^B$. It is easy to see that

$$\begin{aligned}\pi(Y^* | \boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, \boldsymbol{Z}_{S^B}) = \sum_{C \in C^*} & \pi(C | \boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, \boldsymbol{Z}_{S^B}) \\ & \times \pi(Y^* | C, \boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, \boldsymbol{Z}_{S^B}) \end{aligned} \tag{2}$$

Two approaches can be used. In the former we assume that the posterior distribution of $C$ is only affected by the key variables and not by the information provided by $Y$ and $Z$'s. Then expression (2) gets transformed into

$$\pi(Y^* | \boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, Z_{S^B}) = \sum_{C \in C^*} \pi(C | \boldsymbol{X}^A, \boldsymbol{X}^B) \pi(Y^* | C, Y_{S^A}, \boldsymbol{Z}_{S^B}).$$

The first term is related to the linkage step; no information coming from the regression analysis is used. The second term should be analysed in detail. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) \in \Theta$. We know that

$$\pi(Y^* | C, Y_{S^A}, Z_{S^B}) = \int_{\Theta} \pi(y^* | \boldsymbol{\theta}, C, Y_{S^A}, Z_{S^B})\, \pi(\boldsymbol{\theta} | C, Y_{S^A}, Z_{S^B}) d\boldsymbol{\theta}. \tag{3}$$

Given the matrix $C$,some pairs $(y_i, z_j')^{(c)}$ are linked; here the exponent $c$ indicates those pairs which are linked given $C$. If we assume a Normal-Inverse Gamma prior for $(\boldsymbol{\beta}, \sigma^2)$, the posterior will also be of the same type. Then we set

$$\sigma^2 \sim IG(t_0/2, d_0/2), \quad \boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\beta_0}, \sigma^2 \boldsymbol{V_0}).$$

The posterior is

$$\sigma^2 | C, y_{S^A}, z_{S^B} \sim IG(t^*(c)/2, d^*(c)/2) \tag{4}$$

and

$$\boldsymbol{\beta}|\sigma^2, C, y_{S^A}, z_{S^B} \sim N(\boldsymbol{\beta^*(c)}, \sigma^2 \boldsymbol{V^*(c)}). \tag{5}$$

In detail,

$$\boldsymbol{V^*(c)} = (\boldsymbol{Z(c)'Z(c)} + \boldsymbol{V_0^{-1}})^{-1}$$

$$\boldsymbol{\beta^*(c)} = \boldsymbol{V^*(c)}(\boldsymbol{Z(c)'y(c)} + \boldsymbol{V_0}^{-1}\boldsymbol{\beta_0})$$

$$t^*(c) = t_0 + \text{card}(S^A)$$

$$d^*(c) = d_0 + (\text{card}(S^A) - k)S^2 + (\hat{\boldsymbol{\beta}}\boldsymbol{(c)} - \boldsymbol{\beta_0})'\boldsymbol{Z(c)'Z(c)V^*(c)V_0}^{-1}(\hat{\boldsymbol{\beta}}\boldsymbol{(c)} - \boldsymbol{\beta_0})$$

where

$$S^2 = \frac{(\boldsymbol{y(c)} - \hat{\boldsymbol{y}}\boldsymbol{(c)})'(\boldsymbol{y(c)} - \hat{\boldsymbol{y}}\boldsymbol{(c)})}{\text{card}(S^A) - k},$$

$\hat{\boldsymbol{\beta}}\boldsymbol{(c)}$ is the conditional maximum likelihood estimate of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{y}}\boldsymbol{(c)} = \boldsymbol{Z(c)}\hat{\boldsymbol{\beta}}\boldsymbol{(c)}$. Note that these expressions are identical to the usual ones in Bayesian linear modelling, except for the fact that the entire structure is conditioned on $C$. Another point to stress is the following: it might happen that the number of matches implied by $C$ is less than the sample size $n_A$. In this case we need to impute a value for the covariate of the units which are not linked.

The posterior distributions (4) and (5) represent the second term of the integrand in (3). We now need to manage the first term; it is useful to split $Y^*$ as

$$Y^* = Y_1{}^* + Y_2{}^*,$$

where

$$Y_1{}^* = \sum_{i \notin S^A, i \in S^B} y_i \quad \text{and} \quad Y_2{}^* = \sum_{i \notin S^A, i \notin S^B} y_i.$$

It is easy to see that

$$Y_1{}^* | \boldsymbol{\theta}, C, y_{S^A}, z_{S^B} \sim N\left( \sum_{i \notin S^A, i \in S^B} \boldsymbol{z_i}'\boldsymbol{\beta}, \sigma^2 \text{card}(i \notin S^A, i \in S^B) \right)$$

and

$$Y_2{}^* | \boldsymbol{\theta}, C, y_{S^A}, z_{S^B} \sim N\left( g_2(C)[\boldsymbol{\mu(z)}]'\boldsymbol{\beta}, \sigma^2 g_2(C) \right),$$

where $g_2(C)$ is the number of pairs which are not observed (i.e. they were not in file $A$, neither in file $B$) and $\boldsymbol{\mu(z)} = (1, \mu(z_2), \mu(z_3), \ldots, \mu(z_k))'$ is the vector of covariate means. These two distributions are independent then we can easily see that $Y^*|\boldsymbol{\theta}, C, y_{S^A}, z_{S^B}$ has a Gaussian distribution with mean

$$\boldsymbol{\lambda} = \sum_{i \notin S^A, i \in S^B} \boldsymbol{z_i}'\boldsymbol{\beta} + g_2(C)[\boldsymbol{\mu(z)}]'\boldsymbol{\beta},$$

and variance

$$\boldsymbol{\omega}^2 = \sigma^2(\text{card}(i \notin S^A, i \in S^B) + g_2(C)).$$

Using standard arguments, $\boldsymbol{\beta}$ can be easily integrated out: So we can write

$$Y^*|\sigma^2, C, y_{S^A}, z_{S^B} \sim N(a, \sigma^2(b + g_2(C) + \text{card}(i \notin S^A, i \in S^B))) \tag{6}$$

where $a = \sum_{i \notin S^A, i \in S^B} \boldsymbol{z_i}'\boldsymbol{\beta^*} + g_2(C)[\boldsymbol{\mu(z)}]'\boldsymbol{\beta^*}$ and $b = \boldsymbol{p}'\boldsymbol{V^*p}$ with $\boldsymbol{p} = g_2(C)$ $\boldsymbol{\mu(z)} + \boldsymbol{\ell}$ and

$$\boldsymbol{\ell} = \left(\text{card}(i \notin S^A, i \in S^B), \sum_{i \notin S^A, i \in S^B} z_{i,2}, \ldots, \sum_{i \notin S^A, i \in S^B} z_{i,k}\right)'.$$

Finally combining the results in (4) and (6) we obtain the predictive distribution of $Y^*$ as

$$Y^*|C, y_{S^A}, z_{S^B} \sim St_1\left(t^*, a, \frac{d^*}{t^*}(b + g_2(C) + \text{card}(i \notin S^A, i \in S^B))\right),$$

that is a scalar Student $t$ r.v. with $t^*$ dgf. The closed form expression for the density of $Y^*$ allows one to easily simulate from it, according to the following steps.

For $g$ in $1, \ldots, G$:

1. simulate $C^{(g)}$ from $\pi(C|\boldsymbol{X}^A, \boldsymbol{X}^B)$ (using the Metropolis-Hastings algorithm described in [6])
2. simulate $Y^{*(g)}$ from $\pi(Y^*|C^{(g)}, Y_{S^A}, Z_{S^B})$.

An alternative approach is obtained by explicitly allowing the regression part of the model to influence the posterior distribution of $C$. In particular the first term of the integrand in (2) can be expressed as

$$\pi(C|\boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, Z_{S^B}) = \int_{\boldsymbol{\theta} \in \Theta} \pi(C, \boldsymbol{\theta}|\boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, Z_{S^B}).$$

Then it's easy to modify the strategy by only changing the first step of the previous algorithm:

```
for $g$ in $1, \dots, G$:
1. simulate $(C^{(g)}, \boldsymbol{\theta}^{(g)})$ from $\pi(C, \boldsymbol{\theta}|\boldsymbol{X}^A, \boldsymbol{X}^B, Y_{S^A}, Z_{S^B})$
2. simulate $Y^{*(g)}$ from $\pi(Y^*|C^{(g)}, Y_{S^A}, Z_{S^B})$.
```

## 4 A Small Scale Simulation Study

A small simulation has been performed: we have chosen a true $C$ $50 \times 48$ matrix with 26 matches. $N$ is equal to 500. We have generated 10 samples of key variables and 10 samples of $Y$ and a scalar $Z$; We have used 5 key variables and 7 categories for each of them. Each simulation is based on 60000 iterations with a burn-in time of 11000. We distinguish two cases and compare them to the posterior distribution of $Y^*$ using the true $C$, which represents our benchmark. Using the first strategy discussed above the posterior distribution of $Y^*$ is quite different from the benchmark, especially in terms of variability. Using the second strategy, the posterior distribution of $Y^*$ is much more similar to the benchmark: this happens because of a feed-back effect which is able to improve both the linkage and the estimation steps.

## References

1. Copas J, Hilton F (1990) Record linkage: statistical models for matching computer records. J R Stat Soc A 153:287–320
2. Lahiri P, Larsen MD (2005) Regression analysis with linked data. J Am Stat Assoc 100:222–230
3. Scheuren F, Winkler WE (1993) Regression analysis of data files that are computer matched. Surv Methodol 19:39–58
4. Scheuren F, Winkler WE (1997) Regression analysis of data files that are computer matched, Part II. Surv Methodol 23:157–165
5. Tancredi A, Liseo B (2011) A hierarchical Bayesian approach to record linkage and population size problems. Ann Appl Stat 5(2B):1553–1585
6. Tancredi A, Liseo B (2015) Regression analysis with linked data: problems and possible solutions. Statistica 75(1):19–35
7. Tancredi A, Méthe MA, Marcoux M, Liseo B (2013) Accounting for matching uncertainty in two stage capture-recapture experiments using photographic measurements of natural marks. Environ Ecol Stat 20(4):647–665

# The Extension of Imprecise Probabilities Based on Generalized Credal Sets

**Andrey G. Bronevich and Igor N. Rozenberg**

**Abstract** In the paper we continue investigations started in the paper presented at ISIPTA'15, where the notions of lower and upper generalized credal sets has been introduced. Generalized credal sets are models of imprecise probabilities, where it is possible to describe contradiction in information, when the avoiding sure loss condition is not satisfied. The paper contains the basic principles of approximate reasoning: models of uncertainty based on upper previsions and generalized credal sets, natural extension, and coherence principles.

## 1 Introduction

The theory of imprecise probabilities [1, 6, 9] allows us to model conflict (randomness) and non-specificity (imprecision) in information and any model of uncertainty can be equivalently represented by the sets of probability measures also called credal sets. But the modeling of contradiction is not possible. Some authors [5, 7] when dealing with contradiction try to correct it returning to imprecise probability model, but this way of processing uncertainty seems to be not general. Meanwhile, in the theory of evidence [2, 3, 8] contradiction can be modeled by assigning positive values to belief functions at empty set. This way of representing contradiction with small changes was adopted in [4] for conjunction of contradictory sources of information, and where generalized lower and upper credal sets are introduced. Lower and upper credal sets are dual concepts that give us the same way of representing uncertainty. A lower generalized credal set consists of belief functions conceived as upper probabilities whose bodies of evidence contain only singletons and certain

A.G. Bronevich (✉)
National Research University Higher School of Economics, Myasnitskaya 20,
101000 Moscow, Russia
e-mail: brone@mail.ru

I.N. Rozenberg
JSC Research, Development and Planning Institute for Railway Information Technology,
Automation and Telecommunication, Orlikov per. 5, building 1, 107996 Moscow, Russia
e-mail: I.Rozenberg@gismps.ru

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_11

event, thus allowing us to model conflict and contradiction in information; while the whole generalized credal set allows us to model also non-specificity. If an upper prevision does not avoid sure loss, then the corresponding usual credal set is empty and the classical theory of imprecise probabilities is not applicable, but based on generalized credal sets we can analyze this information.

The paper has the following structure. We remind first some definitions from the theory of imprecise probabilities, monotone measures and belief functions. After that we describe how the conjunctive rule can be applied to probability measures and by this way we introduce generalized credal sets. The next sections of the paper are devoted to the main constructions: approximate reasoning based on generalized credal sets, natural extension, and coherence principles.

## 2 The Main Definitions and Constructions

Let $X$ be a finite non-empty set and let $2^X$ be the powerset of $X$. A set function $\mu : 2^X \to [0, 1]$ is called a *monotone measure* if $\mu(\emptyset) = 0$, $\mu(X) = 1$, and $A \subseteq B$ for $A, B \in 2^X$ implies $\mu(A) \leqslant \mu(B)$. On the set of all monotone measures denoted by $M_{mon}$ we introduce the following operations:

- *convex sum*: $\mu = a\mu_1 + (1 - a)\mu_2$ for $\mu_1, \mu_2 \in M_{mon}$ and $a \in [0, 1]$ if $\mu(A) = a\mu_1(A) + (1 - a)\mu_2(A)$ for all $A \in 2^X$;
- *order relation*: $\mu_1 \leqslant \mu_2$ for $\mu_1, \mu_2 \in M_{mon}$ if $\mu_1(A) \leqslant \mu_2(A)$ for all $A \in 2^X$;
- *dual relation*: $\nu = \mu^d$ for $\mu \in M_{mon}$ if $\nu(A) = 1 - \mu(\bar{A})$, where $A \in 2^X$ and $\bar{A}$ is the complement of $A$.

We use also the following constructions from theory of belief functions:

- $Bel : 2^X \to [0, 1]$ is called a *belief function* if $Bel$ can be represented as $Bel(A) = \sum_{B \subseteq A} m(B)$, where $m$ is non-negative set function with $\sum_{B \in 2^X} m(B) = 1$ called the *basic belief assignment* (bba). It is possible that $m(\emptyset) > 0$, when we model contradiction by belief functions. The set $A \in 2^X$ is called a *focal element* for a belief function $Bel$ with bba $m$ if $m(A) > 0$ and set of all focal elements is called the *body of evidence*;
- a belief function is called *categorical* if the body of evidence contains only one focal element $B \in 2^X$. This set function is denoted by $\eta_{\langle B \rangle}$ and can be computed as $\eta_{\langle B \rangle}(A) = 1$ if $B \subseteq A$ and $\eta_{\langle B \rangle}(A) = 0$ otherwise. Every belief function can be represented as $Bel = \sum_{B \in 2^X} m(B)\eta_{\langle B \rangle}$, where $m$ is its bba;
- a belief function is a *probability measure* if its body of evidence consists of singletons, or equivalently $P \in M_{mon}$ is a probability measure if $P(A) + P(B) = P(A \cup B)$ for disjoint sets $A, B \in 2^X$. The set of all probability measures on $2^X$ is denoted by $M_{pr}$.

## 3 Modeling Uncertainty by Imprecise Probabilities

A $\mu \in M_{mon}$ is called a *lower probability* if its values can be viewed as lower bounds of probabilities. This $\mu$ *avoids sure loss* or it is *non-contradictory* if there is $P \in M_{pr}$ such that $\mu \leqslant P$. It is possible to describe uncertainty by lower previsions that can be viewed as lower bounds of mean values of random variables. Let $K$ be the set of all functions $f : X \to \mathbb{R}$ and $K' \subseteq K$. A mapping $\underline{E} : K' \to \mathbb{R}$ can be considered as a *lower prevision functional* if $\underline{E}(f) \geqslant \inf_{x\in X} f(x)$ for all $f \in K'$. For any $P \in M_{pr}$ and $f \in K$ we define the mean value as $E_P(f) = \sum_{x\in X} f(x)P(\{x\})$. Then a lower prevision functional is *non-contradictory* or *avoids sure loss* if the set of probability measures

$$\mathbf{P} = \left\{P \in M_{pr} | \forall f \in K' : \underline{E}(f) \leqslant E_P(f)\right\} \tag{1}$$

is not empty. Let $X = \{x_1, \ldots, x_n\}$, then every $P \in M_{pr}$ can be viewed as a point $(P(\{x_1\}), \ldots, P(\{x_n\}))$ in $\mathbb{R}^n$. By definition, a non-empty set of probability measures is called a *credal set* if it is convex and closed. Clearly, any lower prevision functional defines the corresponding credal set by formula (1) and it is possible to prove that models of imprecise probabilities based on credal sets and lower previsions are equivalent, i.e. every credal set can be generated by a lower prevision. It is easy to see that lower probabilities can be modeled by lower previsions if we define $K' = \{1_A\}_{A\in 2^X}$, where $1_A$ is the characteristic function of the set $A \in 2^X$.

The basic instrument of approximate reasoning in the theory of imprecise probabilities is the natural extension. Let $\underline{E} : K' \to \mathbb{R}$ be a non-contradictory lower prevision functional, then the *natural extension* of $\underline{E}$ is defined as $\underline{E}_{\mathbf{P}}(f) = \inf\{E_P(f) | P \in \mathbf{P}\}$, where $f \in K$ and $\mathbf{P}$ is defined by formula (1). $\underline{E}$ is called a *coherent lower prevision* if $\underline{E}_{\mathbf{P}}(f) = \underline{E}(f)$, $f \in K'$.

If we work with upper bounds of probabilities, then we consider monotone measures called *upper probabilities*. An upper probability $\mu \in M_{mon}$ is *non-contradictory* or it *avoids sure loss* if there is $P \in M_{pr}$ such that $P \leqslant \mu$. Analogously, we introduce an *upper prevision functional* $\bar{E} : K' \to \mathbb{R}$ for $K' \subseteq K$ such that $\underline{E}(f) \leqslant \sup_{x\in X} f(x)$ for all $f \in K'$. It is not contradictory if it defines the credal set

$$\mathbf{P} = \left\{P \in M_{pr} | \forall f \in K' : \bar{E}_P(f) \leqslant \bar{E}(f)\right\}.$$

The natural extension of $\bar{E}$ based on the corresponding credal set $\mathbf{P}$ is defined by $\bar{E}_{\mathbf{P}}(f) = \sup\{E_P(f) | P \in \mathbf{P}\}$. It easy to see that models of uncertainty based on lower and upper previsions are equivalent, we can change lower previsions to upper previsions using the formula: $\bar{E}(f) = -\underline{E}(-f)$, $f \in K'$, because functionals $\underline{E}$ and $\bar{E}$ in this case define the same credal set. This equality holds also for coherent lower and upper previsions: $\bar{E}_{\mathbf{P}}(f) = -\underline{E}_{\mathbf{P}}(-f)$, $f \in K$. If we work with lower probabilities, then we get upper probabilities using the dual relation, i.e. if $\mu$ is a lower probability, then $\mu^d$ is an upper probability.

Let $M \subseteq M_{mon}$, then we denote $M^d = \{\mu^d | \mu \in M\}$. For example, let $M_{bel}$ be the set of all belief functions on $2^X$, then $M^d_{bel}$ is the set of all plausibility functions on $2^X$.

## 4 The Conjunctive Rule for Contradictory Sources of Information Based on Generalized Credal Sets

Let sources of information be described by credal sets $\mathbf{P}_1, \dots, \mathbf{P}_m$. Then if these sources of information are fully reliable, we can use the *conjunctive rule* (C-rule) to aggregate them defined by $\mathbf{P} = \mathbf{P}_1 \cap \dots \cap \mathbf{P}_m$. Let us notice that this rule is applicable if the set $\mathbf{P}$ is not empty. If sources of information are described by lower previsions $\underline{E}_i : K' \to \mathbb{R}$, $i = 1, \dots, m$, then the equivalent C-rule is based on maximum operation: $\underline{E}(f) = \max_i \underline{E}_i(f)$, and the equivalent C-rule for upper previsions is based on minimum operation. Meanwhile, in the theory of belief functions the conjunctive rule can be defined also for contradictory sources of information described by belief functions. For belief functions the C-rule is not defined uniquely, and in [2, 3] the choice based on optimality criteria is analyzed. Based on other grounds, the conjunction of $P_1, \dots, P_m \in M_{pr}$ has been established in [4] as $P = P_1 \wedge \dots \wedge P_m$, where

$$P = a_0 \eta^d_{\langle X \rangle} + \sum_{x_i \in X} a_i \eta_{\langle \{x_i\} \rangle}, \tag{2}$$

is viewed as lower probability and $a_i = \min_j P_j(\{x_i\})$, $i = 1, \dots, n$, and $a_0 = 1 - \sum_{i=1}^n a_i$. The value $Con(P) = a_0$ is called the *measure of contradiction*. Let us denote the set of all monotone measures of the type (2) by $M_{cpr}$.

The C-rule has the following interpretation [4]. If we consider the set $M_{cpr}$ as a partially ordered set w.r.t. $\leqslant$, then $P$ can be viewed as the exact upper bound of the set $\{P_1, \dots, P_m\}$. The last result simply follows from the fact that $P_1 \leqslant P_2$, where $P_1 = a_0 \eta^d_{\langle X \rangle} + \sum_{x_i \in X} a_i \eta_{\langle \{x_i\} \rangle}$ and $P_2 = b_0 \eta^d_{\langle X \rangle} + \sum_{x_i \in X} b_i \eta_{\langle \{x_i\} \rangle}$ iff $a_i \geqslant b_i$, $i = 1, \dots, n$. This allows us to introduce the following definition.

**Definition 1** A subset $\mathbf{P} \subseteq M_{cpr}$ is called an *upper generalized credal set* (UG-credal set) if

(a) $P_1 \in \mathbf{P}$, $P_2 \in M_{cpr}$, $P_1 \leqslant P_2$ implies that $P_2 \in \mathbf{P}$;
(b) if $P_1, P_2 \in \mathbf{P}$, then $aP_1 + (1-a)P_2 \in \mathbf{P}$ for any $P_1, P_2 \in \mathbf{P}$ and $a \in [0, 1]$;
(c) the set $\mathbf{P}$ is closed as a subset of $\mathbb{R}^n$ (any $P$ of the type (2) is a point $(a_1, \dots, a_n)$ in $\mathbb{R}^n$).

Let us describe how we identify usual credal sets in the whole family of UG-credal sets. Let $\mathbf{P}$ be an UG-credal set. Then the set of all minimal elements in $\mathbf{P}$ is called the profile of $\mathbf{P}$ and denoted by *profile*($\mathbf{P}$). Clearly, the *profile*($\mathbf{P}$) defines uniquely the corresponding UG-credal set $\mathbf{P}$. As we will see later if *profile*($\mathbf{P}$) is the usual credal set, i.e. *profile*($\mathbf{P}$) $\subseteq M_{pr}$, then $\mathbf{P}$ brings the same information as the credal set *profile*($\mathbf{P}$). We define the C-rule for UG-sets in the same way as for usual credal sets, i.e. if $\mathbf{P}_1, \dots, \mathbf{P}_m$ are credal sets in $M_{cpr}$, then the C-rule produces the credal set

$$\mathbf{P} = \mathbf{P}_1 \cap \dots \cap \mathbf{P}_m. \tag{3}$$

This definition is justified by the fact that if profiles of $\mathbf{P}_i$ in formula (3) are usual probability measures, i.e. $\mathit{profile}(\mathbf{P}_i) = \{P_i\}$, where $P_i \in M_{pr}$, $i = 1, \dots, m$, then $\mathit{profile}(\mathbf{P}) = \{P\}$ and $P = P_1 \wedge \dots \wedge P_m$.

Further we will use also the dual concept to UG-credal set called the *lower generalized credal set* (LG-credal set), i.e. if $\mathbf{P}$ is an UG-credal set in $M_{cpr}$, then $\mathbf{P}^d$ is the LG-credal set in $M_{cpr}^d$.

*Remark 1* Let $\mathbf{P}$ be an UG-credal set in $M_{cpr}$, then any $P \in \mathbf{P}$ is viewed as lower probability. Conversely, any $P^d \in \mathbf{P}^d$ is viewed as an upper probability and this measure is represented as $P^d = a_0\eta_{\langle X\rangle} + \sum_{i=1}^n a_i\eta_{\langle\{x_i\}\rangle}$. The measure of contradiction of $P^d$ is $Con(P^d) = a_0$. We can reformulate in this way other concepts introduced for UG-credal sets. For example, any $P^d \in M_{cpr}^d$ can be represented as a point $(a_1, \dots, a_n)$ in $\mathbb{R}^n$. Then any credal set $\mathbf{P}^d$ is a convex closed set in $M_{cpr}^d$, but the property a) from Definition 1 should be reformulated as $P_1 \in \mathbf{P}^d$, $P_2 \in M_{cpr}^d$, $P_2 \leqslant P_1$ implies that $P_2 \in \mathbf{P}$. The profile of $\mathbf{P}^d$ consists of all maximal elements in $\mathbf{P}^d$ w.r.t. $\leqslant$ and obviously $\mathit{profile}(\mathbf{P}^d) = (\mathit{profile}(\mathbf{P}))^d$.

## 5 Expectation Estimation with Generalized Credal Sets

Let us analyze how the functional $E_P$ w.r.t. $P \in M_{pr}$ can be extended to monotone measures in $M_{cpr}$ viewed as lower probabilities. Let us look on the C-rule. Let $P \in M_{cpr}$ be the result of aggregating probability measures $P_i$ by the conjunctive rule, then every $P_i \in M_{pr}$ participating in aggregation should be lower than $P$, i.e. $P_i \leqslant P$ and you can find also that $P = \bigwedge_{P_i | P_i \leqslant P} P_i$. Then taking in account that for finding conjunction of lower previsions the maximum operation is used, we come to the formula

$$\underline{E}_P(f) = \sup\left\{E_{P_i}(f) | P_i \leqslant P, P_i \in M_{pr}\right\},$$

where $\underline{E}_P(f)$ can be understood as a lower expectation of $f \in K$ w.r.t. $P \in M_{cpr}$. Because $P \in M_{bel}^d$, we can use for computing $\underline{E}_P(f)$ the Choquet integral, that in this special case gives us the value

$$\underline{E}_P(f) = \sum_{i=1}^n a_i f(x_i) + a_0 \max_{i=1,\dots,n} f(x_i),$$

where $P$ is defined by formula (2). Thus, the lower bound of expectation w.r.t. an UG-credal set $\mathbf{P}$ can be defined by $\underline{E}_{\mathbf{P}}(f) = \inf_{P \in \mathbf{P}} \underline{E}_P(f)$, $f \in K$.

The same constructions can be introduced for LG-credal sets. Assume that $P = a_0\eta_{\langle X\rangle} + \sum\limits_{i=1}^n a_i\eta_{\langle\{x_i\}\rangle}$ in $M_{cpr}^d$ is viewed as a upper probability, then the upper bound of expectation is defined by

$$\bar{E}_P(f) = \sum_{i=1}^{n} a_i f(x_i) + a_0 \min_{i=1,\dots,n} f(x_i),\ f \in K.$$

Analogously, the upper bound of expectation w.r.t. a LG-credal set $\mathbf{P}$ is defined as $\bar{E}_{\mathbf{P}}(f) = \sup_{P \in \mathbf{P}} \bar{E}_P(f),\ f \in K$. The duality relation between functionals $\underline{E}_{\mathbf{P}}$ and $\bar{E}_{\mathbf{P}^d}$ for a UG-credal set $\mathbf{P}$ in $M_{cpr}$ looks the same as for usual credal sets: $\bar{E}_{\mathbf{P}^d}(f) = -\underline{E}_{\mathbf{P}}(-f),\ f \in K$. Therefore, we will formulate next results for LG-credal sets.

The next theorem proved in [4] describes the main properties of the functional $\bar{E}_{\mathbf{P}}$. The non-negative function $f \in K$ is called *normalized* if there is $x_i \in X$, where $f(x_i) = 0$.

**Theorem 1** *A functional $\Phi : K \to \mathbb{R}$ coincides with $\bar{E}_{\mathbf{P}}$ on $K$ for some credal set* $\mathbf{P}$ *in $M_{cpr}^d$ iff it has the following properties:*

*(1) $\Phi(f + a1_X) = \Phi(f) + a$ for any $f \in K$ and $a \in \mathbb{R}$;*
*(2) $\Phi(af) = a\Phi(f)$ for any $f \in K$ and $a \geqslant 0$;*
*(3) $\Phi(f_1) \leqslant \Phi(f_2)$ for $f_1, f_2 \in K$ if $f_1 \leqslant f_2$;*
*(4) $\Phi(f_1) + \Phi(f_2) \geqslant \Phi(f_3)$ for any normalized functions $f_1, f_2, f_3$ in $K$ such that $f_1 + f_2 = f_3$.*

*Remark 2* Let us observe that if we take $a = 0$ in 2) (Theorem 1) we get $\Phi(\mathbf{0}) = 0$, where $\mathbf{0}$ is a function that identical to zero. Putting $f = \mathbf{0}$ and $a = 1$ in 2), we get that $\Phi(1_X) = 1$. The analogous theorem takes place and for the functional $\bar{E}_{\mathbf{P}}$, where $\mathbf{P}$ is a usual credal set, but in that theorem the subadditivity property 4) is fulfilled for arbitrary functions $f_1, f_2, f_3$ in $K$.

It easy to show that there are different LG-credal sets in $M_{cpr}^d$that produce the same generalized upper prevision functional. This can be shown by the following example.

*Example 1* Assume that $X = \{x_1, x_2\}$ and LG-credal sets $\mathbf{P}_1$ and $\mathbf{P}_2$ in $M_{cpr}^d$ are given by their profiles: *profile*$(\mathbf{P}_1) = \{P_1\}$ and *profile*$(\mathbf{P}_2) = \{aP_2 + (1-a)P_3 | a \in [0, 1]\}$, where $P_1 = 0.5\eta_{\langle\{x_1\}\rangle} + 0.5\eta_{\langle\{x_2\}\rangle}$, $P_2 = 0.5\eta_{\langle\{x_1\}\rangle} + 0.5\eta_{\langle X\rangle}$, $P_3 = 0.5\eta_{\langle\{x_2\}\rangle} + 0.5\eta_{\langle X\rangle}$. Then, $\bar{E}_{\mathbf{P}_1}(f) = 0.5f(x_1) + 0.5f(x_2)$ and $\bar{E}_{\mathbf{P}_2}(f) = 0.5\max\{f(x_1), f(x_2)\} + 0.5\min\{f(x_1), f(x_2)\} = \bar{E}_{\mathbf{P}_1}(f)$.

We will try next to describe the case when generalized upper prevision functionals coincide for different LG-sets in $M_{cpr}^d$. For this purpose we will introduce the following definition. Let a LG-credal set $\mathbf{P}$ be described by a convex set in $\mathbb{R}^n$, i.e. we assume that $(a_1, \dots, a_n) \in \mathbf{P}$ if $P = a_0\eta_{\langle X\rangle} + \sum_{i=1}^{n} a_i \eta_{\langle\{x_i\}\rangle}$ is in $\mathbf{P}$. The $j$th projection of $P = (a_1, \dots, a_n)$ in $M_{cpr}^d$ is the point $(b_1, \dots, b_n)$ such that $b_i = a_i$ for all $i \neq j$ and $b_i = 0$ for $i = j$. Obviously, the point $(b_1, \dots, b_n)$ can be interpreted as an element in $M_{cpr}^d$ denoted by $\mathrm{Pr}_j\, P$. The $j$th projection of a LG-credal set $\mathbf{P}$ in $M_{cpr}^d$ is defined as $\mathrm{Pr}_j\, \mathbf{P} = \left\{\mathrm{Pr}_j\, P | P \in \mathbf{P}\right\}$. Clearly, $\mathrm{Pr}_j\, \mathbf{P}$ is also a LG-credal set in $M_{cpr}^d$ and $\mathrm{Pr}_j\, \mathbf{P} \subseteq \mathbf{P}$.

**Theorem 2** *Let* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *be LG-credal sets in* $M_{cpr}^d$*. Then* $\bar{E}_{\mathbf{P}_1}(f) = \bar{E}_{\mathbf{P}_2}(f)$ *for all* $f \in K$ *iff* $\mathrm{Pr}_j\,\mathbf{P}_1 = \mathrm{Pr}_j\,\mathbf{P}_2$, $j = 1, \ldots, n$.

Thus, we see that the functional $\bar{E}_{\mathbf{P}}(f)$ does not define the underlying LG-credal set $\mathbf{P}$ in $M_{cpr}^d$ uniquely. For this reason, let us introduce the following definition.

**Definition 2** The LG-credal set $\mathbf{P}$ in $M_{cpr}^d$ is called *maximal* if

$$\mathbf{P} = \left\{P \in M_{cpr}^d | \forall f \in K : \bar{E}_P(f) \leqslant \bar{E}_{\mathbf{P}}(f) \right\}.$$

**Theorem 3** *Let* $\mathbf{P}$ *be a LG-credal set in* $M_{cpr}^d$ *whose profile is an usual credal set in* $M_{pr}$*. Then the credal set* $\mathbf{P}$ *is maximal.*

## 6 The Natural Extension Based on Generalized Credal Sets

Let $\bar{E} : K' \to \mathbb{R}$ be an upper prevision functional, then it defines the LG-credal set $\mathbf{P} = \left\{P \in M_{cpr}^d | \forall f \in K : \bar{E}_P(f) \leqslant \bar{E}(f)\right\}$ and the functional $\bar{E}_{\mathbf{P}}$ can be considered as the natural extension of $\bar{E}$ on $K$. The functional $\bar{E}$ is called the *generalized coherent prevision* if $\bar{E}_{\mathbf{P}}(f) = \bar{E}(f)$, $f \in K'$.

**Theorem 4** *Let* $\bar{E} : K' \to \mathbb{R}$ *be an upper prevision functional. Then its natural extension* $\bar{E}' : K \to \mathbb{R}$ *based on LG-credal sets can be computed as*

$$\bar{E}'\left(\underline{f}\right) = \inf\left\{\sum_k a_k \bar{E}\left(\underline{f_k}\right) + a | \sum_k a_k \underline{f_k} + a \geqslant \underline{f}, f_k \in K', a_k, a \geqslant 0\right\}, \quad (4)$$

*where* $\underline{f}$, $\underline{f_k}$ *are normalized functions and* $\bar{E}_{\mathbf{P}}\left(\underline{f}\right) = \bar{E}_{\mathbf{P}}(f) - b$, $\bar{E}\left(\underline{f_k}\right) = \bar{E}(f_k) - b_k$, $b = \min_{x \in X} f(x)$, $b_k = \min_{x \in X} f_k(x)$.

*Remark 3* It is easy to see the difference between the natural extensions based on usual credal sets and LG-credal sets. For computing the natural extension based on usual credal sets it is sufficient to allow $a$ to be any real number in formula (4).

*Example 2* Assume that we have two sources of information that describe possible diseases of a patient, and the set of diseases is $X = \{x_1, x_2, x_3\}$. The first source of information certifies that probabilities of events $\{x_1, x_2\}$ and $\{x_2, x_3\}$ are lower or equal to 0.5. The second source of information fully supports that it is disease $x_2$, i.e. the probability of the event $\{x_1, x_3\}$ is equal to 0. If we describe the first source of information by upper previsions, then $K' = \left\{1_{\{x_1,x_2\}}, 1_{\{x_2,x_3\}}\right\}$, and $\bar{E}(1_{\{x_1,x_2\}}) = \bar{E}(1_{\{x_2,x_3\}}) = 0.5$. Consider natural extensions $\bar{E}'$ and $\bar{E}''$ of $\bar{E}$ based on LG-credal sets and usual credal sets. We see that $\bar{E}''(1_{\{x_2\}}) = 0$, i.e. the natural extension based on usual credal sets says that it is definitely not the disease $x_2$, but $\bar{E}'(1_{\{x_2\}}) = 0.5$.

It easy to see that the underlying credal set for $\bar{E}''$ consists of one probability measure $P_1 = (0.5, 0, 0.5)$, and the underlying LG-credal set for $\bar{E}'$ has the profile $\{tP_1 + (1-t)P_2 | t \in [0, 1]\}$, where $P_2 = (0, 0.5, 0)$. We see that sources of information are fully contradictory if we consider usual credal sets and no conclusion can be done. But we can aggregate sources of information using the C-rule and get the LG-credal set with the profile $\{P_2\}$ and make the conclusion that it is disease $x_2$ but with contradiction $Con(P_2) = 0.5$.

## References

1. Augustin T, Coolen FPA, de Cooman G, Troffaes MCM (eds) (2014) Introduction to imprecise probabilities. Wiley, New York
2. Bronevich AG, Rozenberg IN (2014) The choice of generalized Dempster-Shafer rules for aggregating belief functions based on imprecision indices. In: Cuzzolin F (ed) Belief functions: theory and applications. Lecture notes in computer science, vol 8764. Springer, Berlin, pp 21–28
3. Bronevich A, Rozenberg I (2015) The choice of generalized Dempster-Shafer rules for aggregating belief functions. Int J Approximate Reasoning 56:122–136
4. Bronevich AG, Rozenberg IN (2015) The generalization of the conjunctive rule for aggregating contradictory sources of information based on generalized credal sets. In: Augustin T, Doria S, Miranda E, Quaeghebeur E (eds) Proceedings of the 9th International symposium on imprecise probability: theories and applications. Aracne Editrice, Rome, pp 67–76
5. Brozzi A, Capotorti A, Vantaggi B (2012) Incoherence correction strategies in statistical matching. Int J Approximate Reasoning 53:1124–1136
6. Klir GJ (2006) Uncertainty and information: foundations of generalized information theory. Wiley-Interscience, Hoboken
7. Quaeghebeur E (2015) Characterizing coherence, correcting incoherence. Int J Approximate Reasoning 56:208–223
8. Smets Ph (2007) Analyzing the combination of conflicting belief functions. Inf Fusion 8:387–412
9. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, London

# A Generalized SMART Fuzzy Disjunction of Volatility Indicators Applied to Option Pricing in a Binomial Model

**Andrea Capotorti and Gianna Figà-Talamanca**

**Abstract** In this paper we extend our previous contributions on the elicitation of the fuzzy volatility membership function in option pricing models. More specifically we generalize the SMART disjunction for a multi-model volatility behavior (Uniform, LogNormal, Gamma, ...) and within a double-source (direct vs. indirect) information set. The whole procedure is then applied to the Cox-Ross-Rubinstein framework for option pricing on the S&P500 Index where the historical volatility, computed from the Index returns' time series, and the VIX Index observed data are respectively considered as the direct and indirect sources of knowledge. A suitable distance among the resulting fuzzy option prices and the market bid-ask spread make us appreciate the proposed procedure against the classical fuzzy mean.

**Keywords** Smart average operators · Fuzzy mean · Merging · Coherent conditional probabilities · Fuzzy option pricing

## 1 Introduction

In previous contributions [1] we introduced a methodology for membership elicitation on the hidden volatility parameter $\sigma$ of a risky asset through both the historical volatility estimator $\widehat{\sigma}$ and the estimator $\nu = \mathrm{VIX}/100$, based on VIX; Our proposal led on the Coletti and Scozzafava [3] interpretation of membership functions as coherent conditional probability assessments, integrated with observational data, expert evaluations and simulation results. Consequently we followed an *hybrid* approach, lying in between deterministic and stochastic volatility models.

The peculiarity of our procedure was to deal with alternative sources of information, though leaving as an open problem the search for a proper *fusion operator*, the

A. Capotorti (✉)
Dipartimento Matematica e Informatica, Università degli Studi di Perugia, Perugia, Italy
e-mail: andrea.capotorti@unipg.it

G. Figà-Talamanca
Dipartimento Economia, Università degli Studi di Perugia, Perugia, Italy
e-mail: gianna.figatalamanca@unipg.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_12

choice of which is heavily context-dependent. Within this framework, fuzzy arithmetic mean is commonly adopted, being it a basic operation for estimation and also a fuzzy set-theoretic connective.

An bridge between the *estimation* and the *fusion* views of merging information is e.g. Yager's intelligent *constrained merging*. In [2], we borrowed from Yager's proposal the motivation of including an intelligent component in the averaging process to address conflicts in the data to be fused, but, contrarily to the original suggestion, without resorting to an exogenous "combinability function". We named the introduced operators as "SMART", not only because *Smart* is a—inflated—synonymous of *intelligent*, but mainly because as an acronym it means "**S**pecific, **M**easurable, **A**chievable, **R**ealistic and **T**ime-related", as our approach aimed to be.

In [2] we introduced two different kinds of fusion operators: one that *disjointly* considers *different distribution models* and an other for merging *conjointly* the values stemming from the *different estimators*.

The difference between the two suggested operators was on the deformation's direction with respect to the fuzzy arithmetic mean: toward canonical *disjunction*, i.e. *max*; or toward canonical *conjunction*, i.e. *min*.

Both the operators were binary, allowing the merging of couples of fuzzy numbers, and hardly generalizable to $n$-ary. This feature is not a limitation for the specific application of the conjunction, as long as two sources of information are considered, $\widehat{\sigma}$ and $\nu = \text{VIX}/100$, while for the disjunction we were lucky to have the two most contradictory results stemming from different simulating models covering the third one. But this was just for chance in the examples analyzed in the quoted paper, hence we need a more general $n$-ary disjunctive merging operator.

As further novelty, we apply the full methodology, stemming from the memberships elicitation and arriving to the fuzzy options pricing and proper comparisons with bid-ask market prices, to the discrete Cox-Ross-Rubinstein (CRR henceforth) binomial market model. Before we proceed with our specific proposal, we briefly review some of the basic notions about fuzzy numbers and their aggregation.

Membership functions $\mu : \mathbb{R} \to [0, 1]$ of the fuzzy set of possible values of a random variable $X$ are usually viewed as either an imprecise value or as a possibility distribution over a set of crisp values. As already stated, thanks to [3] we can view them as conditional probabilities with the conditioning event varying.

Anyhow, operationally, we will profit from membership characterization through $\alpha$-cuts $\mu^\alpha = \{x \in \mathbb{R} : \mu(x) \geq \alpha\}, \alpha \in [0, 1]$. The $\alpha$ value can be conveniently interpreted as 1 minus the lower bound of the probability that quantity $X$ hits $\mu^\alpha$.

In [1] we were able to elicit membership functions through probability-possibility transformations induced by confidence intervals around the median of specific simulating distributions and we obtained so called "fuzzy numbers", i.e. unimodal membership functions with nested $\alpha$-cuts identified by an interval $[\mu_l^\alpha, \mu_r^\alpha]$ in the extended reals $\widetilde{\mathbb{R}}$. Aggregations are performed between $\alpha$-cuts by considering full/partial overlapping: for the conjunctive operator, we keep the smart $\overline{\wedge}$ computed level-wise in [2] since, as already mentioned in the introduction, we deal with two parameter estimators; we get

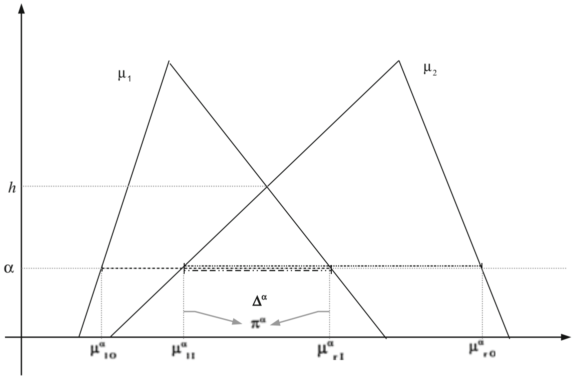


**Fig. 1** Characteristic values for the conjunction of two memberships level-wise

$$
\begin{aligned}
(\mu 1 \barwedge \mu 2)^\alpha &= [(\mu 1 \barwedge \mu 2)^\alpha_l, (\mu 1 \barwedge \mu 2)^\alpha_r] \\
&= \left[ wl^\alpha \mu^\alpha_{lI} + (1 - wl^\alpha)\mu^\alpha_{lO} \ , \ wr^\alpha \mu^\alpha_{rI} + (1 - wr^\alpha)\mu^\alpha_{rO} \right].
\end{aligned}
\tag{1}
$$

where the subscript O refers to the “outer” values, while the subscript I to the “inner” ones (see e.g. Fig. 1). There is a “shrinking” towards the “inner” part for $\alpha$-cuts with non empty intersection—i.e. $\alpha \le h$ in Fig. 1—while towards the “outer” part otherwise—i.e. $\alpha > h$ in Fig. 1. Such shrinking is realized by a careful choice of the weights $wl^\alpha$ and $wr^\alpha$ that are proportional to $\delta^\alpha$ for $\alpha \le h$, while they resume to the arithmetic mean, with a further quadratic deformation, for $\alpha > h$ (for further details refer to the quoted paper).

On the contrary, since the usual simulating models for the volatility are more than two, e.g., Uniform, LogNormal and Gamma, for the disjunctive operator we now propose a new smart merging, again by considering full/partial overlapping of $n$ $\alpha$-cuts. This can be obtained by adapting Marzullo’s algorithm [6], originally designed to compute “relaxed” intersections among different information sources, computing specific weights $\pi^j_f$ and representing the partial overlapping among different $f$ $\alpha$-cuts, $f = 1, \dots, n$. Due to the lack of space we just give here an idea of such quantities by showing them in Fig. 2 and by giving a pseudo code of an R algorithm to compute them in Table 1. Extremes of the $\alpha$-cuts of the disjunctive operator $[(\mu 1 \veebar \dots \veebar \mu n)^\alpha_l, (\mu 1 \veebar \dots \veebar \mu n)^\alpha_r]$, are again computed as convex combinations of the original ones, with $n - 1$ coefficients

$$
\frac{1}{n}(1 + \epsilon^\alpha_j) \qquad j = 1, \dots, n-1 \ , \tag{2}
$$

where the $\epsilon^\alpha_j = \frac{\sum_{f=1}^n \frac{1}{f}\pi^j_f}{\Delta_\alpha}$, with $\Delta_\alpha = \max\{\mu i^\alpha_r\}^n_{i=1} - \min\{\mu i^\alpha_l\}^n_{i=1}$, display the weighted contributions of the $n - 1$ more relevant extremes, i.e. the first $n - 1$ outer ones. Obviously, the $n$-th coefficient, associated to the inner extreme, must be

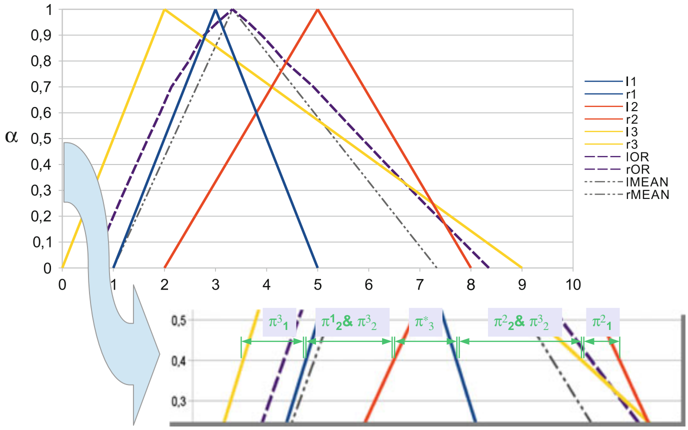


**Fig. 2** SMART disjunction (*dashed line*) among 3 fuzzy numbers compared to the fuzzy arithmetic mean (*dashed-dotted line*). The zoom shows the relaxed intersections computed through adapted Marzullo's algorithm

$$\frac{1}{n}(1 - \sum_{j=1}^{n-1} \epsilon_j^{\alpha}). \tag{3}$$

## 2 Facing the Practical Problem

We go back to the original practical problem of the implicit assessment of fuzzy volatility based on *two different estimators* $\widehat{\sigma}$ and $\nu$, and on three different simulating models *Uniform, LogNormal, Gamma* for the parameter $\sigma$ of interest.

In particular, for each estimator, *different scenarios* are considered on the base of historical data and experts evaluations.

For each scenario and for both estimators it is possible to build *pseudo-memberships* by coherent extension of a-priori information and likelihood values stemming from classical CRR binomial model with the value of $\sigma$ obtained by a random generation from a specific simulating distribution. Parameters of such distributions are computed according to scenarios peculiarities. The empirical values obtained for the estimators permit the selection of most plausible scenarios with associated membership functions for the fuzzy value of the parameter. It is worth noticing that such fuzzy numbers are single whenever there is sure dominance of one scenario over the others, or more than one whenever dominance is partial.

As an illustrative example we can show how our weighted averaging operators work with a multi-period binomial tree with $N = 10$ periods. As for the historical volatility, by observing $\widehat{\sigma}_{obs} = 0.16$, the Log-Normal simulating model furnishes two undominated scenarios, the "medium" and the "high", while the other two models

**Table 1** Pseudo R code of an adapted Marzullo's algorithm, with input ext = list of left and right extremes, i = list of memberships belongings, type = −1 if left ext; +1 otherwise

```
relaxint = function(ext,i,type)
{
n = length(ext)/2
pi = matrix(0,n,n)
lambda = numeric(2*n)
lambda[1] = - type[1]
j=list()
j[[1]] = c(i[1])
   for (l in 1:(2*n-1))
   { for (k in j[[l]]){
   pi[lambda[l],k] = pi[lambda[l],k] + (ext[l+1] - ext[l])}
    lambda[l+1] = lambda[l] - type[l+1]
    if (type[l+1] == -1)  j[[l+1]]=c(j[[l]], i[l+1])
    else j[[l+1]] = setdiff(j[[l]], c(i[l+1]))
  }
return(pi)
}
```

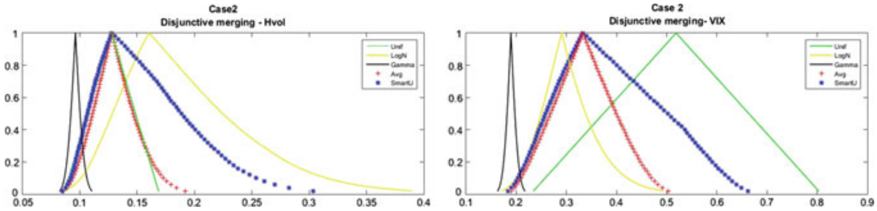


**Fig. 3** Fuzzy estimations of the parameters $\widehat{\sigma}$ (*left*) and $\nu$ (*right*) obtained through the three different simulating model and their merging through $\veebar$ and arithmetic mean

(the Uniform and the Gamma) agree in selecting only the "medium" one, whose associated fuzzy numbers for $\widehat{\sigma}$ are reported in Fig. 3 (left—where the membership associated to the Log-Normal model is already the smart disjunction of the fuzzy numbers stemming from the "medium" and the "high" scenarios). In respect of the other estimator $\nu$, its observed value $\nu_{obs} = 0.19$ always leads to the selection of the "high" scenario, obtaining for its fuzzy estimation the three memberships reported in Fig. 3 (right). We can finally merge with our smart $\veebar$ operator the memberships of each estimator and average the two through $\overline{\wedge}$, obtaining as final unique fuzzy number $\mu_{\widehat{\sigma}_{obs}} \overline{\wedge}\, \mu_{\nu_{obs}}$ whose membership is reported in Fig. 4. Once a fuzzy number

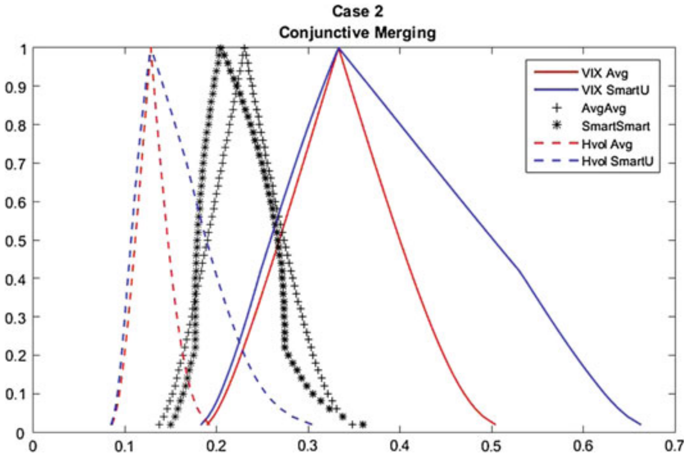


**Fig. 4** The merging results: disjunction of the fuzzy numbers stemming from different models for $\widehat{\sigma}$ (*dashed lines* on the *left*) and for $\nu$ (*solid lines* on the *right*) and their final conjunction (*starred lines* on the *center*), by applying our $\veebar$ and $\overline{\wedge}$ or the arithmetic mean

elicitation for $\tilde{\sigma}$ is obtained through the merging procedure described in the previous section, it is possible to price options by a straightforward extension of standard CRR to fuzzy multi-period binomial model. Our explicit numerical evaluation of each $\alpha$-cut of the fuzzy number for $\tilde{\sigma}$ allows us to take advantage of a different contribution available in literature for each step of the pricing procedure. In particular:

- from $\tilde{\sigma}$ to the the fuzzy "UP" and "DOWN" jump factors (*Zadeh's extension principle* [8])

$$[\underline{u}^{\alpha}, \overline{u}^{\alpha}] = [e^{\underline{\sigma}^{\alpha}\sqrt{\Delta t}}, e^{\overline{\sigma}^{\alpha}\sqrt{\Delta t}}] \qquad [\underline{d}^{\alpha}, \overline{d}^{\alpha}] = [e^{-\overline{\sigma}^{\alpha}\sqrt{\Delta t}}, e^{-\underline{\sigma}^{\alpha}\sqrt{\Delta t}}]\,; \tag{4}$$

- $\tilde{u}$ and $\tilde{d}$ to the fuzzy risk neutral probabilities (Muzzioli and Torricelli [7])

$$[\underline{p}_u^{\alpha}, \overline{p}_u^{\alpha}] = \left[\frac{e^{r\Delta t} - \overline{d}^{\alpha}}{\overline{u}^{\alpha} - \overline{d}^{\alpha}}, \frac{e^{r\Delta t} - \underline{d}^{\alpha}}{\underline{u}^{\alpha} - \underline{d}^{\alpha}}\right] [\underline{p}_d^{\alpha}, \overline{p}_d^{\alpha}] = \left[\frac{\underline{u}^{\alpha} - e^{r\Delta t}}{\underline{u}^{\alpha} - \underline{d}^{\alpha}}, \frac{\overline{u}^{\alpha} - e^{r\Delta t}}{\overline{u}^{\alpha} - \overline{d}^{\alpha}}\right]\,; \tag{5}$$

- $\tilde{p}_u$ and $\tilde{p}_d$ to option price (e.g. call) (Li and Han [5])

$$[\underline{C}_0^{\alpha}, \overline{C}_0^{\alpha}] = e^{-rN\Delta t}\left[\sum_{i=0}^{N}(\underline{p}_u^{\alpha})^i(\underline{p}_d^{\alpha})^{N-i}\underline{C}_{N,i}^{\alpha}, \sum_{i=0}^{N}(\overline{p}_u^{\alpha})^i(\overline{p}_d^{\alpha})^{N-i}\overline{C}_{N,i}^{\alpha}\right] \tag{6}$$

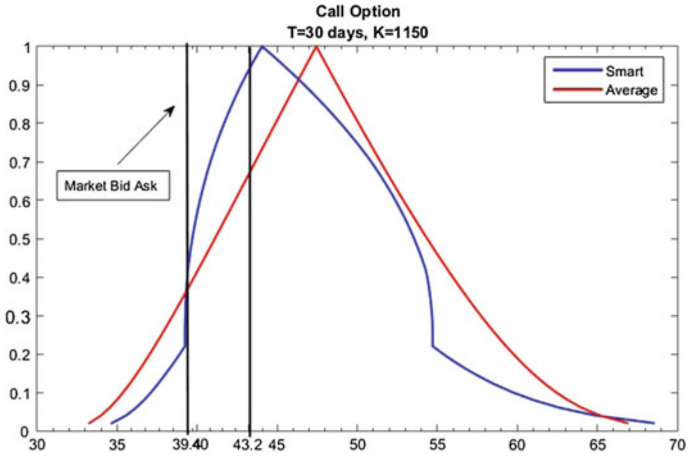


**Fig. 5** Fuzzy option prices: market bid-ask (crisp interval), "smart" (*blue*), "arithmetic mean" (*red*)

with

$$[\underline{C}^{\alpha}_{N,i}, \overline{C}^{\alpha}_{N,i}] = \left[\max(S_0(\underline{u}^{\alpha})^i(\underline{d}^{\alpha})^{N-i} - K, 0), \max(S_0(\overline{u}^{\alpha})^i(\overline{d}^{\alpha})^{N-i} - K, 0)\right]. \tag{7}$$

According to the fuzzy number obtained by suitably merging information on volatility, we compute the corresponding fuzzy option prices for SPX options written on the S&P500 Index on a specific date.[1]

In order to appreciate the capability of our procedure to capture market option prices, the Bhattacharya distance

$$R(A, B) = \left[1 - \int_{-\infty}^{+\infty} (\mu^*_A(x)\mu^*_B(x))^{1/2} dx\right]^{1/2} \tag{8}$$

with $\mu^*_{\cdot}(x) = \mu_{\cdot}(x)/\text{Power}(\cdot)$ and $\text{Power}(\cdot) = \int_{-\infty}^{+\infty} \mu_{\cdot}(x)dx$ is computed between fuzzy model prices and the corresponding market bid-ask prices thought as *crisp intervals*; this is done both for our *"smart"* model fuzzy prices and for those derived by the *fuzzy arithmetic mean* (see an example in Fig. 5). The outcomes seem to be deeply influenced by the a priori choice of the number of volatility's scenarios and their possible overlapping. In fact, if we refer to the different scenarios detection proposed in [1], by choosing just three incompatible scenarios like in Case 1 distances of our fuzzy option prices from bid-ask interval are sensibly worst than those obtained with usual arithmetic mean. On the contrary, with three partially overlapping scenarios like in Case 2 our method performs slightly better than arithmetic mean if we

[1] Reported results refer to trading on October 21st, 2010.

consider all kinds of options (51 %) or only those ”near the money”—NDM—(52 %), and sensibly better for those ”at the money”—ATM—(60 %).

Much better performances are obtained with the more fine five scenarios partition of Case 3, where our price are closer to the market in 81 % of all the traded options, that became 86 % if we focus only on the NTM and we reach 96 % by considering just those ATM.

## 3 Conclusion

We proposed a complete procedure for computing fuzzy option prices in the CRR environment. Starting from the volatility membership elicitation (usually assumed as known), based on a multi-model (Uniform, LogNormal, Gamma) volatility behavior and with a double-source (direct v indirect) information set, and thanks to original smart merging operators $\underline{\vee}$ and $\overline{\wedge}$, the suggested methodology performs quite well by comparing model prices to market bid-ask prices via a fuzzy-distance measure. Further efforts are in order to possibly define a better similarity measure—e.g. by weighting differently the values $x$ in the Power$(\cdot)$ function—apt to capture the closeness between fuzzy prices and crisp bid-ask intervals, and to define a reasonable merging of conjunctive fusion levels among $n > 2$ sources.

**Acknowledgments** This work was partially supported by INdAM-GNAMPA through the Project 2015 U2015/000418 and by DMI—Universitá degli Studi di Perugia ricerca di base project 2016 “Operatori SMART per aggregazioni fuzzy”.

## References

1. Capotorti A, Figá-Talamanca G (2013) On an implicit assessment of fuzzy volatility in the Black and Scholes environment. Fuzzy Sets Syst 223:59–71
2. Capotorti A, Figá-Talamanca G (2014) Smart fuzzy weighted averages of information elicited through fuzzy numbers. In: Laurent A et al (eds) Information processing and management of uncertainty in knowledge-based systems. IPMU 2014, Part I, CCIS 442, pp 466–475
3. Coletti G, Scozzafava R (2004) Conditional probability, fuzzy sets, and possibility: a unifying view. Fuzzy Sets Syst 144:227–249
4. Dubois D, Prade H, Yager RR (1999) Merging fuzzy information. In: Bezdek JC, Dubois D, Prade H (eds) Fuzzy Sets in approximate reasoning and information systems. The Handbooks of Fuzzy Sets Series. Kluwer Academic Publishers, Dordrecht, pp 335–401
5. Li W, Han L (2009) The fuzzy binomial option pricing model under Knightian uncertainty. In: 2009 Sixth International conference on fuzzy systems and knowledge discovery
6. Marzullo K (1990) Tolerating failures of continuous-valued sensors. ACM Trans Comput Syst 8(4):284–304
7. Muzzioli S, Torricelli C (2004) A multiperiod binomial model for pricing options in a vague world. J Econ Dyn Control 28:861–887
8. Zadeh LA (1975) The concept of linguistic variable and its application to approximate reasoning. Inf Sci 8:199–249

# The Representation of Conglomerative Functionals

**Gianluca Cassese**

**Abstract** We prove results concerning the representation of certain linear functionals based on the notion of conglomerability, originally introduced by Dubins and de Finetti. We show that this property has some applications in probability and in statistics.

## 1 Introduction

Take the sets $\Omega$, $\Omega'$ and $S$ and the family $\mathscr{H} \subset \mathbb{R}^S$ as given. Fix a map $X \in S^\Omega$ and a finitely additive probability $m$ on $\Omega$ with $h(X) \in L^1(m)$. Hereafter we study the problem of finding $X' \in S^{\Omega'}$ and $\mu$ on $\Omega'$ such that

$$\int h(X)dm = \int h(X')d\mu \qquad h \in \mathscr{H}. \tag{1}$$

In the terminology of Dubins and Savage [3], $X$ and $X'$ are then *companions*, a property depending on $\mathscr{H}$, our model for the information available.

With $\Omega = S = \mathbb{R}$ and $X$ the identity, the left hand side of (1) may, e.g., originate from some experiment for which a correct mathematical model $X'$ is sought and each $h \in \mathscr{H}$ is a statistic. A similar problem arises in Bayesian statistics: given a predictive marginal $m$ the purpose is to find a parametric family $\{Q_\theta : \theta \in \Theta\}$ of probabilities and a prior $\lambda$ on $\Theta$ such that

$$m(A) = \int_\Theta Q_\theta(A)d\lambda \tag{2}$$

A problem similar to (2) was investigated long ago by Dubins [4] in terms of a special condition, conglomerability, originally due to de Finetti. In this work we provide a

G. Cassese (✉)
Dipartimento di Economia, Metodi Quantitativi E Strategie D'Impresa,
Universitá Milano Bicocca, via Bicocca Degli Arcimboldi 8, 20126 Milan, Italy
e-mail: gianluca.cassese@unimib.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_13

general formulation of such property and show some of its possible applications to probability and statistics.

In particular we obtain in Theorem 1, our main result, an integral representation for conglomerative linear functionals on an arbitrary vector space. Despite its simplicity, this result admits a large number of implications of which the existence of a solution to (1) is just a case in point.

## 2 Notation and Preliminary Results

Throughout the paper $\mathscr{A}$ and $\Sigma$ are rings of subsets of $\Omega$ and $S$, respectively. $\mathfrak{L}(\Omega, S)$ and $\mathscr{C}(\Omega, S)$ denote the families of linear and continuous maps $f \in S^\Omega$, respectively (reference to $S$ is omitted when $S = \mathbb{R}$). If $f \in S^\Omega$ and $A \subset \Omega$ the image of $A$ under $f$ is indicated as $f[A]$. A sublattice $\mathscr{H}$ of $\mathbb{R}^S$ is Stonean, if $h \in \mathscr{H}$ implies $h \wedge 1 \in \mathscr{H}$.

$\mathscr{S}(\mathscr{A})$ is the family of $\mathscr{A}$ simple functions on $\Omega$ and $\mathfrak{B}(\mathscr{A})$ its closure in the topology of uniform convergence. By $fa(\mathscr{A})$ and $ba(\mathscr{A})$ we designate the spaces of real valued, finitely additive set functions on $\mathscr{A}$ and those elements of $fa(\mathscr{A})$ which are bounded in variation, respectively. If $\mathscr{A}$ is a ring of subsets of $\Omega$ and $\lambda \in fa(\mathscr{A})_+$ the pair $(\mathscr{A}, \lambda)$ is a measurable structure.

If $\lambda \in fa(\mathscr{A})_+$ we say that $X \in \mathbb{R}^\Omega$ is $\lambda$-measurable if and only if there exists a sequence $\langle X_n \rangle_{n\in\mathbb{N}}$ in $\mathscr{S}(\mathscr{A})$ that $\lambda$-converges to $X$, i.e. such that $\lim_n \lambda^*(|X_n - X| > c) = 0$ for every $c > 0$ where $\lambda^*$ and its conjugate $\lambda_*$ are defined (with the convention $\inf \varnothing = \infty$) as $\lambda^*(E) = \inf_{\{A\in\mathscr{A}:E\subset A\}} \lambda(A)$ and $\lambda_*(E) = \sup_{\{B\in\mathscr{A}:B\subset E\}} \lambda(B)$ for all $E \subset \Omega$. $X$ is $\lambda$-integrable, $X \in L^1(\lambda)$, if there is a sequence $\langle X_n \rangle_{n\in\mathbb{N}}$ in $\mathscr{S}(\mathscr{A})$ that $\lambda$-converges to $X$ and is Cauchy in $L^1(\lambda)$. If $A, B \subset \Omega$ and $f \in L^1(\lambda)$, then

$$\mathbb{1}_A \le f \le \mathbb{1}_B \quad \text{implies} \quad \lambda^*(A) \le \int f d\lambda \le \lambda_*(B). \tag{3}$$

Associated with $\lambda \in fa(\mathscr{A})_+$ and $X \in \mathbb{R}^\Omega$ are the following collections:

$$D(X, \lambda) = \left\{t \in \mathbb{R} : \lim_n \lambda_*(X > t - 2^{-n}) = \lim_n \lambda_*(X > t + 2^{-n})\right\} \tag{4a}$$

$$\mathscr{R}_0(X, \lambda) = \left\{\{X > t\} : t \in D(X, \lambda)\right\} \tag{4b}$$

$$\mathscr{A}(\lambda) = \left\{E \subset \Omega : \lambda^*(E) = \lambda_*(E) < \infty\right\}. \tag{4c}$$

$\lambda$ admits but one extension to $\mathscr{A}(\lambda)$ and $X$ is $\lambda$-measurable if and only if it is measurable with respect to such extension, denoted again by $\lambda$. A sequence $\langle X_n \rangle_{n\in\mathbb{N}}$ in $L^1(\lambda)$ converges to $X$ in norm if and only if it is Cauchy in norm and $\lambda$-converges to $X$. If $S$ is a topological space, then $X \in S^\Omega$ is $\lambda$-tight if for all $\varepsilon > 0$ there exists $K \subset S$ compact such that $\lambda^*(X \notin K) < \varepsilon$.

We now state without proof some basic facts concerning finite additivity, some of which well known under countable additivity. We fix $\lambda \in fa(\mathscr{A})_+$.

**Lemma 1** *$X \in \mathbb{R}^\Omega$ is $\lambda$-measurable if and only if it is $\lambda$-tight and either (i) $\lambda_*(X > s) \geq \lambda^*(X \geq t)$ for all $s < t$ or (ii) $\mathscr{R}_0(X, \lambda) \subset \mathscr{A}(\lambda)$.*

**Lemma 2** *If $X \in L^1(\lambda)$ then $\int X d\lambda = \int_0^\infty \lambda_*(X > t)dt - \int_{-\infty}^0 \lambda_*(X < \tau)d\tau$.*

To prove uniqueness of the set function generating a given class of integrals we need to identify a minimal measurable structure. More precisely we define an order $\preceq$ for measurable structures on the same underlying space by writing

$$(\mathscr{A}, \lambda) \preceq (\mathscr{B}, \xi) \quad \text{whenever} \quad \mathscr{A} \subset \mathscr{B}(\xi) \quad \text{and} \quad \xi|\mathscr{A} = \lambda. \tag{5}$$

Speaking of a minimal measurable structure refers to such partial order.

**Lemma 3** *Let $\mathscr{H} \subset \mathbb{R}^\Omega$ be a Stonean vector lattice and $\phi \in \mathfrak{L}(\mathscr{H})_+$. The family $\mathfrak{M}(\phi)$ of measurable structures $(\mathscr{A}, \lambda)$ on $\Omega$ satisfying*

$$\mathscr{H} \subset L^1(\lambda) \quad \textit{and} \quad \int h d\lambda = \phi(h) \qquad h \in \mathscr{H}, \tag{6}$$

*is either empty or contains a minimal element $(\mathscr{R}_\phi, \lambda_\phi)$.*

*Proof* Let $(\mathscr{A}, \lambda) \in \mathfrak{M}(\phi)$ and let $\mathscr{R}_\phi$ be the smallest ring containing

$$\mathscr{R}_0(\phi) = \big\{\{h > t\} : h \in \mathscr{H}_+,\ t \in D(h, \lambda),\ t > 0\big\}. \tag{7}$$

Write $\lambda_\phi = \lambda|\mathscr{R}_\phi$. $\varnothing \in \mathscr{R}_0(\phi)$, as $\mathscr{H}$ is Stonean. Suppose that $(\mathscr{B}, \xi) \in \mathfrak{M}(\phi)$. Fix $h \in \mathscr{H}_+$ and consider the classical inequality

$$\mathbb{1}_{\{h>a\}} \geq (h \wedge b - h \wedge a)/(b - a) \geq \mathbb{1}_{\{h \geq b\}} \qquad h \in \mathscr{H}, b > a > 0. \tag{8}$$

By the Stone property, the inner term belongs to $\mathscr{H}$, so that $\infty > \lambda_*(h > a) \geq \xi^*(h \geq b)$, by (3). Choosing $a$ and $b$ conveniently and interchanging $\lambda$ with $\xi$ we establish that $D(h, \lambda) \cap (0, \infty) = D(h, \xi) \cap (0, \infty)$ and that

$$\lambda^*(h \geq t) = \xi^*(h \geq t) = \xi_*(h > t) = \lambda_*(h > t) \qquad t \in D(h, \lambda),\ t > 0$$

so that $\mathscr{R}_0(\phi) \subset \mathscr{B}(\xi)$. For $i = 1, 2$ pick $h_i \in \mathscr{H}_+$, $t_i \in D(h_i, \lambda)$ and $t_i > 0$. One can easily prove that $\mathscr{R}_0(\phi)$ is closed with respect to union and intersection. Because $\lambda$ and $\xi$ are additive and coincide on $\mathscr{R}_0(\phi)$ they also coincide on $\mathscr{R}_\phi$, [1, Theorem 3.5.1]. Let $h \in \mathscr{H}_+$ and $t > s$. Then $h$ is $\lambda_{\mathscr{H}}$-tight because $h \in L^1(\lambda)$. If $s < 0$ then $\lambda_{\phi*}(h > s) \geq \lambda_\phi^*(h \geq t)$. Otherwise there are $t', s' \in D(h, \lambda)$ with $t > t' > s' > s$ and therefore $\lambda_{\phi*}(h > s) \geq \lambda_\phi(h > s') \geq \lambda_\phi(h > t') \geq \lambda_\phi^*(h \geq t)$. By Lemma 1 $h$ is thus $\lambda_\phi$ -measurable and therefore $\int h d\lambda_\phi = \int h d\lambda$.

The minimal measurable $(\mathscr{R}_\phi, \lambda_\phi)$ will generally depend on $\lambda$. However, since $D(h,\lambda)$ is dense, the generated $\sigma$ ring corresponds to the usual notion.

**Lemma 4** *Let $g : \Omega \to \mathbb{R}_+$ be $\lambda$-measurable and define the ring $\mathscr{R}_g = \big\{A \in \mathscr{A}(\lambda) : g\mathbb{1}_A \in L^1(\lambda)\big\}$. There exists a unique $\lambda_g \in fa(\mathscr{R}_g)_+$ such that*

$$\int f\lambda_g = \int fgd\lambda \qquad f \in \mathfrak{B}(\lambda),\ fg \in L^1(\lambda). \tag{9}$$

The identity on $\Omega' = S$ is a companion to whatever random quantity $X$.

**Proposition 1** *Let $m \in fa(\mathscr{A})_+$, $X \in S^\Omega$. Let $\mathscr{H} \subset \mathbb{R}^S$ be a Stonean vector lattice. There is a minimal measurable structure $(\mathscr{R}, \mu)$ on $X[\Omega]$ with*

$$h \in L^1(\mu) \quad and \quad \int h(X)dm = \int hd\mu \qquad h \in \mathscr{H},\ h(X) \in L^1(m). \tag{10}$$

*$\mu$ is countably additive whenever: (i) $m$ is countably additive or (ii) $\mathscr{H} \subset \mathscr{C}(S)$ and either (a) $X$ is $m$-tight or (b) each $h \in \mathscr{H}$ has compact support.*

*Proof* The existence is easily proved by letting

$$\bar{\mathscr{R}} = \big\{B \subset X[\Omega] : X^{-1}(B) \in \mathscr{R}(\mathscr{H}[X], m)\big\} \tag{11}$$

and $\bar{\mu}(B) = m(X \in B)$ for all $B \in \bar{\mathscr{R}}$. Then $(\bar{\mathscr{R}}, \bar{\mu})$ is a measurable structure on $X[\Omega]$ and $D(h(X), m) = D(h, \bar{\mu})$ for every $h \in \mathscr{H}$. By Lemma 1, $h$ is $\bar{\mu}$-measurable; by Lemma 2 $\int h(X)dm = \int hd\bar{\mu}$ so that $\phi(h) = \int hd\bar{\mu}$. By Lemma 3 there is a minimal measurable structure with this property.

$\phi$ is a Daniell integral when $m$ is countably additive. To prove the same under $(ii)$ we follow Karandikar [6] quite closely. We only need to consider case $(a)$, as the restriction to compact sets is obvious under $(b)$. Let the sequence $\langle h_k \rangle_{k\in\mathbb{N}}$ in $\mathscr{H}$ decrease to 0. For each $n \in \mathbb{N}$, let $A_n \in \mathscr{A}$, $A_n \subset \{X \in K_n\}$ and $m(A_n^c) < 2^{-n}$, for some $K_n \subset S$ compact. Then, $h_k(X)\mathbb{1}_{A_n}$ converges uniformly to 0 and, by [5, III.2.15],

$$\lim_k \int h_k(X)dm = \lim_k \lim_n \int h_k(X)\mathbb{1}_{A_n}dm = \lim_n \lim_k \int h_k(X)\mathbb{1}_{A_n}dm = 0$$

which proves the Daniell property.

Claim $(ii)$ was originally formulated, with $S = \Omega = \mathbb{R}$, by Dubins and Savage [3, p.190]; Karandikar [6] revived and extended their proof.

## 3 Integral Representation of Linear Functionals

We prove a theorem on the integral representation of linear functionals.

**Theorem 1** *Let $\mathscr{H}$ be a vector space and $\phi \in \mathfrak{L}(\mathscr{H})$. Assume that $T \in \mathfrak{L}\big(\mathscr{H}, \mathbb{R}^{\Omega}\big)$ satisfies*

$$\forall h \in \mathscr{H},\ \exists h' \in \mathscr{H} \quad \text{such that} \quad |Th| \le Th' \tag{12}$$

*and write $L = \big\{f : \Omega \to \mathbb{R} : |f| \le Th \text{ for some } h \in \mathscr{H}\big\}$. The condition*

$$\phi(h) < 0 \quad \text{implies} \quad \inf_{\omega}(Th)(\omega) < 0 \qquad h \in \mathscr{H} \tag{13}$$

*is necessary and sufficient for the existence of (i) $F^{\perp} \in \mathfrak{L}(L)_{+}$ with $F^{\perp}[L \cap \mathfrak{B}(\Omega)] = \{0\}$ and (ii) a measurable structure $(\mathscr{R}, \mu)$ on $\Omega$ such that $L \subset L^1(\mu)$ and*

$$\phi(h) = F^{\perp}\big(Th\big) + \int Th d\mu \qquad h \in \mathscr{H}. \tag{14}$$

*Proof* By (13), we can define $F \in \mathfrak{L}(T[\mathscr{H}])_{+}$ by letting

$$F\big(Th\big) = \phi(h) \qquad h \in \mathscr{H} \tag{15}$$

and then extend $F$ as a positive linear functional (still denoted by $F$) on $L$. For each $\alpha \subset \mathscr{H}$ finite, let $h_\alpha \in \mathscr{H}$ be such that $Th_\alpha \ge \bigvee_{h \in \alpha} |Th|$, $\Omega_\alpha = \{Th_\alpha \ne 0\}$ and define $I_\alpha(f)(\omega) = f(\omega)/Th_\alpha(\omega)$ when $f \in L$ and $\omega \in \Omega_\alpha$. Let also $L_\alpha = \{f \in L : |f| \le c\, Th_\alpha \text{ for some } c > 0\}$ and $H_\alpha = I_\alpha[L_\alpha]$. Upon writing $U_\alpha\big(I_\alpha(f)\big) = F(f)$ whenever $f \in L_\alpha$ we obtain another positive, linear functional $U_\alpha$ on $H_\alpha$. [2, Theorem1] implies

$$U_\alpha\big(I_\alpha(f)\big) = \int I_\alpha(f) d\bar{m}_\alpha \qquad f \in L_\alpha \tag{16}$$

for some $\bar{m}_\alpha \in ba(\Omega_\alpha)_{+}$. Let $m_\alpha(A) = \bar{m}_\alpha(A \cap \Omega_\alpha)$ for each $A \subset \Omega$. By Lemma 4, we can write (with the convention $0/0 = 0$)

$$F(f) = \int \mathbb{1}_{\Omega_\alpha} f/Th_\alpha dm_\alpha = \int f d\bar{\mu}_\alpha \qquad f \in L_\alpha \cap \mathfrak{B}(\Omega) \tag{17}$$

with $\bar{\mu}_\alpha = m_{\alpha,g}$ defined as in (9) with $g = \mathbb{1}_{\Omega_\alpha}/Th_\alpha$. Since $L_\alpha \cap \mathfrak{B}(\Omega)$ is Stonean, there is by Lemma 3 a minimal $(\mathscr{R}_\alpha, \mu_\alpha)$ supporting (17). Define $\mathscr{R} = \bigcup_\alpha \mathscr{R}_\alpha$ and $\mu(A) = \lim_\alpha \mu_\alpha(A)$ for all $A \in \mathscr{R}$. $\alpha \subset \alpha'$ implies $L_\alpha \subset L_{\alpha'}$, $(\mathscr{R}_\alpha, \mu_\alpha) \preceq (\mathscr{R}_{\alpha'}, \mu_{\alpha'})$ and $\mu_\alpha = \mu_{\alpha'}|\mathscr{R}_\alpha = \mu|\mathscr{R}_\alpha$. If $f \in L_\alpha$ and $f \ge 0$,

$$\begin{aligned} F(f) &= \lim_k F(f \wedge k) + \lim_k F((f-k)^{+}) = \lim_k \int (f \wedge k) d\mu + F^{\perp}(f) \\ &= \int f d\mu + F^{\perp}(f) \end{aligned} \tag{18}$$

where $F^{\perp}(f) = \lim_k F((f-k)^{+})$ and the inequality $\mu^*(f > k) \le k^{-1} F(f)$ implies that $f \wedge k$ is $\mu$-convergent to $f$ and is Cauchy in $L^1(\mu)$. $\int |f| d\mu \le F(|f|)$

follows from (18) and implies $L \subset L^1(\mu)$. (14) is a consequence of (15) and (18). Necessity is obvious as the right hand side of (14) defines a positive linear functional on $L$.

(12) asserts that the range $T[\mathscr{H}]$ of $T$ is an upward directed set with respect to the natural order of $\mathbb{R}^\Omega$ and for this reason we shall refer to it to by saying that $T$ is *directed*. A positive linear map on an upward directed vector space is directed. An immediate corollary is the following representation of positive linear functionals on vector lattices that may fail to be Stonean.

(13) simply requires that $\phi$ and $T$ do not rank the elements of $\mathscr{H}$ in a totally opposite way. We shall refer to it by saying that $\phi$ is $T$-*conglomerative*. To make the connection with the work of Dubins [4] more transparent we establish the following version of the problem considered by him:

**Corollary 1** *Let $(\mathscr{B}, \lambda)$ be a measurable structure on $S \times \Omega$ and $\mathscr{H}$ a Stonean sublattice of $L^1(\lambda)$. Let $\{\sigma_\omega : \omega \in \Omega\} \subset \mathfrak{L}(\mathscr{H})_+$. The condition*

$$\int h d\lambda < 0 \quad \textit{implies} \quad \inf_\omega \sigma_\omega(h) < 0 \qquad h \in \mathscr{H} \tag{19}$$

*is equivalent to the existence of a measurable structure $(\mathscr{R}, \gamma)$ on $\Omega$ such that*

$$\int h d\lambda = \int \sigma_\omega(h) d\gamma \qquad h \in \mathscr{H}. \tag{20}$$

*Proof* Apply Theorem 1 with $T : \mathscr{H} \to \mathbb{R}^\Omega$ defined as $Th(\omega) = \sigma_\omega(h)$.

In [4], $\mathscr{H} = \mathfrak{B}(S)$ and $\Omega$ is a partition of $S$, the family $\sigma$ is called a *strategy* and a probability such as $\lambda$ is called *strategic*.

As for the Bayesian problem in the Introduction, we easily obtain:

**Corollary 2** *Let $m$ and $\{Q_\theta : \theta \in \Theta\}$ be probabilities on $\Sigma$. Write $F_A(\theta) = Q_\theta(A)$ for $A \in \Sigma$. There is a probability $\lambda$ on $\Theta$ such that*

$$F_A \in L^1(\lambda) \quad \textit{and} \quad m(A) = \int Q_\theta(A) d\lambda \qquad A \in \Sigma \tag{21}$$

*if and only if the following condition holds*

$$\int h dm < 0 \quad \textit{implies} \quad \inf_{\theta \in \Theta} \int h dQ_\theta < 0 \qquad h \in \mathscr{S}(\Sigma). \tag{22}$$

## 4 Finitely Additive Representations

**Theorem 2** *Let $m \in fa(\Sigma)_+$, $\mathscr{H}$ a Stonean vector sublattice of $L^1(m)$ and $X' \in S^{\Omega'}$. There is equivalence between the condition*

$$\int hdm < 0 \quad implies \quad \inf_{\omega} h\big(X'(\omega)\big) < 0 \qquad h \in \mathscr{H} \tag{23}$$

*and the existence of a minimal measurable structure* $(\mathscr{R}, \mu)$ *on* $\Omega'$ *satisfying*

$$h(X') \in L^1(\mu) \quad and \quad \int hdm = \int h(X')d\mu \qquad h \in \mathscr{H}. \tag{24}$$

*Either one of* (23) *or* (24) *is implicit in* $m^*(X'[\Omega']^c) = 0$. *If* $m$ *is countably additive,* $\Sigma$ *a* $\sigma$ *ring and* $m_*(X'[\Omega']^c) = 0$ *then* $\mu$ *is countably additive.*

*Proof* (23) is equivalent to (13) with $\phi(h) = \int hdm$ and $Th = h(X')$ for $h \in \mathscr{H}$. Thus, (24) follows from (14) after noting that, in the present setting, $\phi(h) = \lim_k \phi(h \wedge k)$ for every $h \in \mathscr{H}_+$. If $\langle B_n \rangle_{n \in \mathbb{N}}$ is a decreasing sequence in $\Sigma$ with $X'[\Omega']^c \subset B_n$ and $m(B_n) \leq 2^{-n}$ and if $h \in \mathscr{H}$ then

$$\int hdm = \lim_n \int h \mathbb{1}_{B_n^c} dm \geq \inf_{\omega' \in \Omega'} h(X'(\omega')) \lim_n m(B_n^c)$$

Let $m$ be countably additive, $\Sigma$ a $\sigma$ ring and $m_*(X'[\Omega']^c) = 0$. Let $\langle h_n \rangle_{n \in \mathbb{N}}$ be a sequence in $\mathscr{H}$ with $h_n(X')$ decreasing to 0. If $g \in \mathscr{H}$ and $t \in D(g, m) \cap D(g(X'), \mu)$, then from (11), $\{g > t\} \in \Sigma(m)$, $\{g(X') > t\} \in \mathscr{R}(\mu)$ and $m(g > t) = \mu(g(X') > t)$. On a dense subset of $t > 0$ the following holds:

$$m(h_{n+1} - h_n > t) = \mu(h_{n+1}(X') - h_n(X') > t) = 0 \qquad n = 1, 2, \ldots$$

$h_n \downarrow h = \inf_n h_n$, $m$ a.s. so that $\{h > \varepsilon\} \subset X'[\Omega']^c$ and $\{h > \varepsilon\} \in \Sigma$. Therefore, $0 = m(h > 0)$ and $\lim_n \int h_n(X')d\mu = \lim_n \int h_n dm = \int hdm \leq 0$.

In the absence of restrictions on $\mu$, the existence of representations is guaranteed by conglomerability. If, e.g., $X'[\Omega]^c \in \Sigma$, then in order for $X'$ to represent $m$ when $\mathscr{H} = L^1(m)$ it is necessary and sufficient that $m(X'[\Omega']^c) = 0$. If $m$ consists of sample frequencies, then this condition means that all the observations in the given sample must belong to the range of $X$.

*Example 1* Let $(\Omega, \mathscr{A}, P)$ be a classical probability space, $S = \mathbb{R}$ and let $X'$ be normally distributed on $\Omega$. Fix $m \in ba(\mathscr{B}(\mathbb{R}))_+$ arbitrarily and let $\mathscr{H} = \mathscr{C}(\mathbb{R}) \cap L^1(m)$. Given that $P(X' \in B) > 0$ for every $B$ open, we conclude that $m$ is $X'$-conglomerative relatively to $\mathscr{H}$.

In Theorem 2 the representing measure $\mu$ is completely unrestricted. A possible mitigation is to require that $\mu$ vanishes on an ideal $\mathscr{N}$ of subsets of $\Omega'$, i.e. $N, M \in \mathscr{N}$ and $A \subset N$ imply $N \cup M, A \in \mathscr{N}$.

**Theorem 3** *If* $m$, $\mathscr{H}$ *and* $X'$ *are as in Theorem 2, then*

$$\int hdm < 0 \quad implies \quad \sup_{N \in \mathscr{N}} \inf_{\omega' \in N^c} h\big(X'(\omega')\big) < 0 \tag{25}$$

*if and only if there is a measurable structure* $(\mathscr{R}, \mu)$ *with* $\mathscr{N} \subset \mathscr{R}$, $\mu[\mathscr{N}] = \{0\}$,

$$h(X') \in L^1(\mu) \quad \textit{and} \quad \int h dm = \int h(X') d\mu \qquad h \in \mathscr{H}. \tag{26}$$

*If* $\Sigma$ *is a* $\sigma$ *ring,* $m$ *is countably additive,* $\mathscr{N}$ *is closed with respect to countable unions and* $m_*\big(X'[N^c]^c\big) = 0$ *for all* $N \in \mathscr{N}$ *then* $\mu$ *is countably additive.*

*Proof* Write $f \succeq g$ if $\sup_{N \in \mathscr{N}} \inf_{\omega' \in N^c} (f - g)(\omega') \geq 0$. $\succeq$ is a partial order, $f \geq g$ implies $f \succeq g$ and $f_i \succeq g_i$ for $i = 1, 2$ implies $f_1 \vee f_2 \succeq g_1 \vee g_2$. In fact, $f_1 \vee f_2 \succeq f_i \succeq g_i$ i.e. $f_1 \vee f_2 \geq g_i - \varepsilon$ outside of some $N_i \in \mathscr{N}$. Thus, $f_1 \vee f_2 \geq g_1 \vee g_2 - \varepsilon$ outside of $N_1 \cup N_2 \in \mathscr{N}$ i.e. to $f_1 \vee f_2 \succeq g_1 \vee g_2$. Relatively to pointwise ordering, the set $\mathscr{F} = \big\{f \in \mathbb{R}^{\Omega'} : f \mathscr{S} h(X') \text{ for some } h \in \mathscr{H}\big\}$ is a Stonean vector lattice. Define $\phi \in \mathfrak{L}(\mathscr{F})_+$ implicitly via

$$\phi(f) = \int h dm \qquad f \mathscr{S} h(X'),\ h \in \mathscr{H} \tag{27}$$

We conclude that there is a minimal measurable structure $(\mathscr{R}, \mu)$ satisfying

$$f \in L^1(\mu) \quad \text{and} \quad \phi(f) = \int f d\mu \qquad f \in \mathscr{F} \tag{28}$$

If $N \in \mathscr{N}$ then $\mathbb{1}_N \sim 0$, $\mathbb{1}_N \in \mathscr{F}$ and $\mu(N) = 0$. (26) holds; the converse is obvious.

Under the stated conditions the functional $\phi$ in (27) is a Daniell integral. Let $f_n \downarrow 0$ in $\mathscr{F}$ with $f_n \mathscr{S} h_n(X')$ and $h_n \in \mathscr{H}$ for $n = 1, 2, \ldots$. Define $g_n = \bigwedge_{1 \leq j \leq n} h_j$ and $g_n = \lim_n g_n$. We saw that $f_n \mathscr{S} g_n(X')$. Of course, $f_n \succeq g(X')$ so that $\{g(X') > \varepsilon\} \subset \bigcup_n \{g(X') \geq f_n + \varepsilon\} \in \mathscr{N}$ and $\{g > \varepsilon\} \subset X[\{g(X') \leq \varepsilon\}]^c$. Then, $m(g > \varepsilon) = 0$ and $\lim_n \int f_n d\mu = \lim_n \int g_n dm = \int g dm = 0$.

The preceding result has immediate implications for Brownian motion.

**Corollary 3** *Let* $X$ *be Brownian motion and let* $(m_{t_1, \ldots, t_n} : t_1, \ldots, t_n \in \mathbb{R}_+)$ *be a projective family with* $m_{t_1, \ldots, t_n} \in fa(\mathscr{B}(\mathbb{R}^n))$. *There exists a probability space* $(\Omega, \mathscr{A}, Q)$ *such that*

$$m_{t_1, \ldots, t_n}(B) = Q(X_{t_1}, \ldots, X_{t_n} \in B) \qquad B \in \mathscr{B}(\mathbb{R}^n). \tag{29}$$

*Proof* Let $\alpha = \{t_1 < \ldots < t_n\} \subset \mathbb{R}_+$ and let $B \subset \mathbb{R}^n$ be open with $\{s_1, \ldots, s_n\} \in B$. Then there exist $B_1, \ldots, B_n \subset \mathbb{R}$ open such that $s_i' - s_{i-1}' \in B_i$ for $i = 1, \ldots, n$ (and $s_0' = 0$) implies $\{s_1', \ldots, s_n'\} \in B$. By the property of normally distributed, independent increments, $P(X_{t_1}, \ldots, X_{t_n} \in B) > 0$. Conglomerability obtains for $h_\alpha \in \mathscr{C}(\mathbb{R}^n)$.

## References

1. Bhaskara Rao KPS, Bhaskara Rao M (1993) Theory of carges. Academic Press, London
2. Cassese G (2009) Sure wins, separating probabilities and the representation of linear functionals. J Math Anal Appl 354:558–563
3. Dubins LE, Savage LJ (1965) How to gamble if you must. McGraw, New York
4. Dubins LE (1975) Finitely additive conditional probability, conglomerability and disintegrations. Ann Probab 3:89–99
5. Dunford N, Schwartz JT (1988) Linear operators. Part I. Wiley, New York
6. Karandikar R (1982) A general principle for limit theorems in finitely additive probability. Trans Am Math Soc 273:541–550

# The Likelihood Interpretation of Fuzzy Data

Marco E.G.V. Cattaneo

**Abstract** The interpretation of degrees of membership as statistical likelihood is probably the oldest interpretation of fuzzy sets. It allows in particular to easily incorporate fuzzy data and fuzzy inferences in statistical methods, and sheds some light on the central role played by extension principle and $\alpha$-cuts in fuzzy set theory.

**Keywords** Fuzzy sets · Foundations · Likelihood function · Measurement error · Fuzzy inference

## 1 Introduction

Most works on fuzzy set theory do not give any precise interpretation for the values of membership functions. This is not a problem as far as the works remain in the realm of pure mathematics. However, as soon as examples of application are included an interpretation is needed, otherwise not only the membership functions are arbitrary, but also all rules applied to them are unjustified [3, 25, 32].

In this paper, the interpretation of the values of membership functions in terms of likelihood is reviewed. The concepts of probability and likelihood were clearly distinguished by Fisher [19]: likelihood is simpler, more intuitive, and better suited to information fusion [6, 8]. The likelihood interpretation of fuzzy sets is elucidated in Sect. 2, while Sect. 3 shows that it justifies an expression for the likelihood function induced by fuzzy data that appeared often in the literature [13, 20, 23, 26, 35], but without a clear justification. This likelihood function can also be interpreted as resulting from an errors-in-variables model or measurement error model [5], as will be illustrated by a simple example. Finally, Sect. 4 discusses the interpretation of $\alpha$-cuts as confidence intervals, while the last section concludes the paper and outlines future work.

M. Cattaneo (✉)
Department of Mathematics, University of Hull, Kingston upon Hull, UK
e-mail: m.cattaneo@hull.ac.uk

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_14

## 2 The Likelihood Interpretation

A fuzzy set is described by its membership function $\mu : \mathcal{X} \to [0, 1]$, where $\mathcal{X}$ is a nonempty (crisp) set [34]. A standard example is the fuzzy set representing the meaning of the word "tall" in relation to a man, where the elements of $\mathcal{X}$ are the possible values of a man's height in cm [36]. We can expect for instance that $\mu(180) > \mu(160)$, because the attribute "tall" fits better to a 180 cm man than to a 160 cm one. However, the concept of a fuzzy set as described by a real-valued membership function $\mu$ can only be used to model the reality if we have an interpretation for the numerical values of $\mu$.

In fact, a clear interpretation of membership functions should be the starting point of a theory of fuzzy sets that describes the real world, and all rules of the theory should be a consequence of the interpretation [3, 25, 32]. This is for example the case with the theory of probability, whose rules are a consequence of each of its interpretations (at least on finite spaces). As suggested by this example, it is not necessary that the interpretation is unique, but only the rules that are implied by the considered interpretation should be used in applications.

One of the first aspects to consider when discussing the interpretation of fuzzy sets is if they are used in an epistemic or ontic sense [13, 15]. Fuzzy sets have an ontic interpretation when they are themselves the object of inquiry, while they have an epistemic interpretation when their membership function $\mu : \mathcal{X} \to [0, 1]$ only gives information about the real object of inquiry, which is the value of $x \in \mathcal{X}$. In this paper, we will only consider epistemic fuzzy sets, and focus on their interpretation in terms of likelihood.

The likelihood interpretation of a fuzzy set consists in interpreting its membership function $\mu : \mathcal{X} \to [0, 1]$ as the likelihood function $lik$ on $\mathcal{X}$ induced by the observation of an event $D$:

$$\mu(x) = lik(x \mid D) \propto P(D \mid x)$$

for all $x \in \mathcal{X}$, where $P(D \mid x)$ was the probability of the event $D$ (before its realization) given the value of $x \in \mathcal{X}$.

For example, "John is tall" is a piece of information that can be modeled by a fuzzy set with membership function $\mu : \mathcal{X} \to [0, 1]$ with $\mu(x) \propto P(D \mid x)$, where the elements of $\mathcal{X}$ are the possible values of John's height in cm, and $P(D \mid x)$ is the probability of the event $D$ of getting the information that "John is tall" when John's height is $x$ cm. Hence, the exact meaning of the interpretation of fuzzy sets in terms of likelihood depends on the interpretation given to probability values, but as noted above, the choice of this interpretation does not affect the rules of probability theory.

The likelihood interpretation is probably the oldest interpretation of fuzzy sets: it has been more or less explicitly used directly after [27] and even before [2, 29] the mathematical concept of fuzzy set was introduced by Zadeh [34], and has later been studied in detail by several authors [1, 10–12, 14, 16, 17, 22, 24, 30, 31]. However, most of them interpreted membership functions $\mu$ in terms of probability values $\mu(x) = P(D \mid x)$, instead of likelihood values $\mu(x) = lik(x \mid D)$. Historically, the subtle distinction between probability and likelihood confused several great minds, before the likelihood of $x \in \mathcal{X}$ was clearly defined by Fisher as *proportional* to the probability of the data $D$ given $x$ [18, 19, 21].

The proportionality constant in the definition of $lik(x \mid D)$ can depend on anything but the value of $x \in \mathcal{X}$. The reason for defining the likelihood function $lik$ only up to a multiplicative constant is that otherwise $lik$ would strongly depend on irrelevant information. For example, if two persons chosen at random from a population independently tell us that John is "tall" and "very tall", respectively, then the resulting fuzzy set should not change completely if we would or would not have the additional information that the first person said "tall" and the second one "very tall".

Interpreting fuzzy sets in terms of likelihood thus implies that proportional membership functions have the same meaning. Uniqueness of representation is recovered by assuming, as is often done anyway, that all fuzzy sets are *normalized*. That is, their membership functions $\mu : \mathcal{X} \to [0, 1]$ satisfy $\sup_{x \in \mathcal{X}} \mu(x) = 1$, and are thus uniquely determined by $\mu(x) \propto P(D \mid x)$. Surprisingly, very few authors seem to have somehow considered this important aspect of the likelihood interpretation, and not in a very explicit way [14, 25, 31].

## 3 Fuzzy Data

A basic advantage of the likelihood interpretation of fuzzy sets is that it allows to directly obtain statistical inferences from fuzzy data. The only condition on the statistical methods used is that the data enter them through the likelihood function only. In particular, all methods from the likelihood and Bayesian approaches to statistics can be straightforwardly generalized to the case of fuzzy data.

As discussed in Sect. 2, the membership function of a fuzzy set $\mu(x) \propto P(D \mid x)$ is interpreted as the likelihood function induced by the observation of an event $D$. Now, if we have a probability distribution on $x \in \mathcal{X}$, depending on an unknown parameter $\theta \in \Theta$, then the observation of the event $D$ induces also a likelihood function $lik$ on $\Theta$:

$$lik(\theta \mid D) \propto P(D \mid \theta) = \int_{\mathcal{X}} P(D \mid x)\, dP(x \mid \theta) \propto \int_{\mathcal{X}} \mu(x)\, dP(x \mid \theta) \qquad (1)$$

for all $\theta \in \Theta$, where $P(D \mid x)$ is assumed to be a measurable function of $x$ that does not depend on $\theta$.

Zadeh [35] defined the probability of the fuzzy event described by a membership function $\mu : \mathcal{X} \to [0, 1]$ as the right-hand side of (1), without justifying this choice through a clear interpretation of the values of $\mu$. The likelihood interpretation provides only a partial justification: the right-hand side of (1) is proportional to the probability of the event $D$ that induced the fuzzy information described by $\mu$, where the proportionality constant can depend on anything but $\theta$ (or $x$).

In [35] Zadeh introduced also the concept of probabilistic independence for fuzzy events, again without a clear justification. The likelihood interpretation clarifies another concept of independence, which is extremely important in fuzzy set theory: the concept of independence among the pieces of information described by different fuzzy sets, which is usually implicitly or explicitly assumed [3, 24]. The pieces of information described by the membership functions $\mu_1, \dots, \mu_n : \mathcal{X} \to [0, 1]$ with $\mu_i(x) \propto P(D_i \mid x)$ can be interpreted as independent when the events $D_1, \dots, D_n$ that induced them were conditionally independent given $x$. In this case, the joint fuzzy information is described by the membership function $\mu : \mathcal{X} \to [0, 1]$ with

$$\mu(x) = lik(x \mid D) \propto P(D \mid x) = \prod_{i=1}^{n} P(D_i \mid x) \propto \prod_{i=1}^{n} \mu_i(x) \tag{2}$$

for all $x \in \mathcal{X}$, where $D = D_1 \cap \cdots \cap D_n$.

In particular, if $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, the components $x_i$ of $x = (x_1, \dots, x_n)$ are probabilistically independent (for all $\theta$), and each piece of fuzzy information $\mu_i(x_i) \propto P(D_i \mid x)$ is about a different component of $x$, then the assumption of their independence is very natural, and by combining (1) and (2) we obtain

$$lik(\theta \mid D) \propto \int_{\mathcal{X}} \prod_{i=1}^{n} \mu_i(x_i)\, dP(x \mid \theta) = \prod_{i=1}^{n} \int_{\mathcal{X}_i} \mu_i(x_i)\, dP(x_i \mid \theta) \tag{3}$$

for all $\theta \in \Theta$. This likelihood function has been considered by several authors [13, 20, 23, 26], but only justified on the basis of Zadeh's rather arbitrary definition of the probability of a fuzzy event [35].

The likelihood function (3) induced by fuzzy data with membership functions $\mu_i : \mathcal{X}_i \to [0, 1]$ is often too complex to be handled analytically [20], but this is nowadays a typical situation in the likelihood and Bayesian approaches to statistics. In particular, $x_1, \dots, x_n$ play the role of unobserved variables in (3), and therefore the EM algorithm can be used to maximize the likelihood [13]. Several examples of numerical calculations of maximum likelihood estimates based on fuzzy data are given for instance in [13, 23].

When the data are fuzzy numbers, in the sense that $\mathcal{X}_i \subseteq \mathbb{R}$, the likelihood function (3) can also be interpreted as resulting from an errors-in-variables model or measurement error model [5]. In this case, the value $\xi_i$ of a proxy $x_i^*$ is assumed to be observed instead of the value of the variable $x_i$, where $\xi_i \in \mathbb{R}$ is an arbitrarily chosen constant, while the measurement error $\varepsilon_i = x_i^* - x_i$ is random with density $f_i \propto \mu_i(\xi_i - \cdot)$ and independent of everything else. In this model, each fuzzy number $\mu_i(x_i) \propto f_i(\xi_i - x_i) \propto lik(x_i \mid x_i^* = \xi_i)$ describes the information about the unknown value of $x_i$ obtained from the observed value of its proxy $x_i^*$, and the likelihood function $lik(\,\cdot \mid x_1^* = \xi_1, \dots, x_n^* = \xi_n)$ on $\Theta$ induced by these observations is the one in (3). The description of fuzzy data in terms of measurement errors is particularly useful when the various components combine well mathematically, as in the following simple example.

*Example 1* Assume that $x_{1,} \ldots , x_n$ is a sample from a normal distribution with known variance $\sigma^2$ and unknown expectation $\theta \in \mathbb{R}$, but we have only fuzzy data with membership functions $\mu_i(x_i) = \exp\left(-{}^{(x_i-\xi_i)^2}/_{(2\,\sigma_i^2)}\right)$, where $\xi_i, \sigma_i$ are known constants. Then the proxy variables $x_1^*, \ldots , x_n^*$ are independent, and each $x_i^*$ is normally distributed with expectation $\theta$ and variance $\sigma^2 + \sigma_i^2$. Hence, the likelihood function induced by the fuzzy data is given by

$$lik(\theta \,|\, x_1^* = \xi_1, \ \ldots, \ x_n^* = \xi_n) \propto \exp\left(-\tfrac{(\theta-\hat{\theta})^2}{2\,\tau^2}\right) \tag{4}$$

for all $\theta \in \mathbb{R}$, where the maximum likelihood estimate $\hat{\theta}$ is the weighted average of the centers $\xi_i$ of the fuzzy numbers, with weights ${}^{\tau^2}/_{(\sigma^2+\sigma_i^2)}$ depending on their precision ${}^{1}/_{\sigma_i^2}$, while ${}^{1}/_{\tau^2} = \sum_{i=1}^{n} {}^{1}/_{(\sigma^2+\sigma_i^2)}$ is the precision of $\hat{\theta}$ (which is normally distributed with expectation $\theta$ and variance $\tau^2$).

Besides the maximum likelihood estimate $\hat{\theta}$, for each $\alpha \in (0, 1)$ we obtain a likelihood-based confidence interval for $\theta$:

$$\left\{\theta \in \mathbb{R} : lik(\theta) > \alpha\, lik(\hat{\theta})\right\} = \left(\hat{\theta} \pm \tau\,\sqrt{-2\,\ln\alpha}\right), \tag{5}$$

with exact level $F_{\chi_1^2}(-2\,\ln\alpha)$, where $F_{\chi_1^2}$ is the cumulative distribution function of the chi-squared distribution with 1 degree of freedom. Alternatively, we can combine the likelihood function (4) induced by the fuzzy data with a Bayesian prior, and base our conclusions on the resulting posterior. In particular, if the prior is a normal distribution with expectation $\theta_0$ and variance $\tau_0^2$, then the posterior is a normal distribution with expectation $\theta_1$ and variance $\tau_1^2$, where $\theta_1$ is the weighted average of $\theta_0$ and $\hat{\theta}$, with weights proportional to their precision ${}^{1}/_{\tau_0^2}$ and ${}^{1}/_{\tau^2}$, respectively, while these add up to the posterior precision ${}^{1}/_{\tau_1^2} = {}^{1}/_{\tau_0^2} + {}^{1}/_{\tau^2}$.

## 4 Fuzzy Inference

Besides allowing the direct use of fuzzy data in statistical methods, the likelihood interpretation of fuzzy sets also leads naturally to fuzzy statistical inference. In fact, the likelihood function on $\Theta$ induced by the (fuzzy or crisp) data can be interpreted as the membership function $\mu : \Theta \to [0, 1]$ of a (normalized) fuzzy set describing the information obtained from the data about the unknown value of the parameter $\theta \in \Theta$.

In particular, the likelihood-based confidence intervals (or regions) for $\theta$, defined as in the left-hand side of (5) for all $\alpha \in (0, 1)$, correspond to the $\alpha$-cuts of the fuzzy set with membership function $\mu$. Both likelihood-based confidence intervals and $\alpha$-cuts are usually defined using the non-strict inequality, but the choice of the strict inequality in the definition provides a better agreement with the concept of profile likelihood function [9], which is of central importance in the likelihood approach to statistics, and corresponds to the extension principle [36], which is equally central in fuzzy set theory.

A correspondence between $\alpha$-cuts and (general) confidence intervals has also been suggested as an alternative interpretation of some fuzzy sets [4, 28]. However, this interpretation is afflicted by the fact that confidence intervals are rather arbitrary constructs, and in particular do not usually satisfy the extension principle, when they are not likelihood-based confidence intervals. The interpretation of fuzzy sets in terms of likelihood-based confidence intervals (i.e. the likelihood interpretation) has the advantage of uniqueness, invariance, and general applicability, although a simple expression for the confidence level based on the chi-squared distribution, as in Example 1, is valid (exactly or asymptotically) only under some regularity conditions [33].

Since each value of $\theta \in \Theta$ corresponds to a probability measure $P(\cdot \mid \theta)$, a fuzzy set with membership function $\mu : \Theta \to [0, 1]$ can also be interpreted as a fuzzy probability measure [6, 7]. This likelihood-based model of fuzzy probability bears important similarities to the Bayesian model of probability, and can be used as a basis for statistical inference and decision making [6–8].

## 5 Conclusion

In this paper, the likelihood interpretation of fuzzy sets has been reviewed and some of its consequences analyzed. Not surprisingly, with this interpretation fuzzy data and fuzzy inferences can be easily incorporated in statistical methods. In particular, the likelihood interpretation of fuzzy data justifies the use of expression (3) for the induced likelihood function, and establishes a fruitful connection with errors-in-variables models or measurement error models, as illustrated by Example 1. Furthermore, the link between this interpretation and the likelihood approach to statistics sheds some light on the central role played by extension principle and $\alpha$-cuts in fuzzy set theory.

The theory of fuzzy sets is also a theory of information fusion. However, only the product rule $\mu(x) \propto \prod_{i=1}^{n} \mu_i(x)$ for the conjunction of independent pieces of information is directly justified by the likelihood interpretation (2). The rules for other logical connectives, with or without the independence assumption, can be obtained through the concept of profile likelihood (i.e. the extension principle). For example, the conjunction without independence assumption is then given by the minimum rule $\mu(x) \propto \bigwedge_{i=1}^{n} \mu_i(x)$, while negation always results in the vacuous membership function $\mu \equiv 1$. Such rules, which are a consequence of the likelihood interpretation of fuzzy sets, will be the topic of future work.

## References

1. Bilgiç T, Türkşen IB (2000) Measurement of membership functions: theoretical and empirical work. In: Prade H, Dubois D (eds) Fundamentals of fuzzy sets. Springer, pp 195–230
2. Black M (1937) Vagueness. Philos Sci 4:427–455
3. Bradley J (2009) Fuzzy logic as a theory of vagueness: 15 conceptual questions. In: Seising R (ed) Views on fuzzy sets and systems from different perspectives. Springer, pp 207–228
4. Buckley JJ (2006) Fuzzy probability and statistics. Springer, New York
5. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models, 2nd edn. Chapman & Hall/CRC
6. Cattaneo M (2008) Fuzzy probabilities based on the likelihood function. In: Lubiano MA, Prade H, Gil MÁ, Grzegorzewski P, Hryniewicz O, Dubois D (eds) Soft methods for handling variability and imprecision. Springer, New York, pp 43–50
7. Cattaneo M (2009) A generalization of credal networks. In: Augustin T, Coolen FPA, Moral S, Troffaes MCM (eds) ISIPTA '09, SIPTA, pp 79–88
8. Cattaneo M (2013) Likelihood decision functions. Electron J Stat 7:2924–2946
9. Cattaneo M, Wiencierz A (2012) Likelihood-based imprecise regression. Int J Approx Reason 53:1137–1154
10. Coletti G, Scozzafava R (2004) Conditional probability, fuzzy sets, and possibility: a unifying view. Fuzzy Sets Syst 144:227–249
11. Coletti G, Vantaggi B (2010) From comparative degrees of belief to conditional measures. In: Squillante M, Yager RR, Kacprzyk J, Greco S, Marques Pereira RA (eds) Preferences and decisions. Springer, pp 69–84
12. Coletti G, Vantaggi B (2013) Inference with probabilistic and fuzzy information. In: Seising R, Trillas E, Moraga C, Termini S (eds) On fuzziness, vol 1. Springer, pp 115–119
13. Denœux T (2011) Maximum likelihood estimation from fuzzy data using the EM algorithm. Fuzzy Sets Syst 183:72–91
14. Dubois D (2006) Possibility theory and statistical reasoning. Comput Stat Data Anal 51:47–69
15. Dubois D, Prade H (2012) Gradualness, uncertainty and bipolarity: making sense of fuzzy sets. Fuzzy Sets Syst 192:3–24
16. Dubois D, Moral S, Prade H (1997) A semantics for possibility theory based on likelihoods. J Math Anal Appl 205:359–380
17. Dubois D, Nguyen HT, Prade H (2000) Possibility theory, probability and fuzzy sets. In: Prade H, Dubois D (eds) Fundamentals of fuzzy sets. Springer, pp 343–438
18. Edwards AWF (1974) The history of likelihood. Int Stat Rev 42:9–15
19. Fisher RA (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1:3–32
20. Gil MÁ, Casals MR (1988) An operative extension of the likelihood ratio test from fuzzy data. Stat Pap 29:191–203
21. Hald A (1999) On the history of maximum likelihood in relation to inverse probability and least squares. Stat Sci 14:214–222
22. Hisdal E (1988) Are grades of membership probabilities? Fuzzy Sets Syst 25:325–348
23. Jung HY, Lee WJ, Yoon JH, Choi SH (2014) Likelihood inference based on fuzzy data in regression model. In: SCIS & ISIS 2014, IEEE, pp 1175–1179
24. Kovalerchuk B (2014) Probabilistic solution of Zadeh's test problems. In: Laurent A, Strauss O, Bouchon-Meunier B, Yager RR (eds) Information processing and management of uncertainty in knowledge-based systems, vol 2. Springer, pp 536–545
25. Lindley DV (2004) Comment to [31]. J Am Stat Assoc 99:877–879
26. Liu X, Li S (2013) Cumulative distribution function estimation with fuzzy data: Some estimators and further problems. In: Berthold MR, Moewes C, Gil MÁ, Grzegorzewski P, Hryniewicz O, Kruse R (eds) Synergies of soft computing and statistics for intelligent data analysis. Springer, pp 83–91
27. Loginov VI (1966) Probability treatment of Zadeh membership functions and their use in pattern recognition. Eng Cybern 4:68–69

28. Mauris G (2008) Inferring a possibility distribution from very few measurements. In: Lubiano MA, Prade H, Gil MÁ, Grzegorzewski P, Hryniewicz O, Dubois D (eds) Soft methods for handling variability and imprecision. Springer, pp 92–99
29. Menger K (1951) Ensembles flous et fonctions aléatoires. C R Acad Sci 232:2001–2003
30. Scozzafava R (2013) The membership of a fuzzy set as coherent conditional probability. In: Seising R, Trillas E, Moraga C, Termini S (eds) On fuzziness, vol 2. Springer, pp 631–635
31. Singpurwalla ND, Booker JM (2004) Membership functions and probability measures of fuzzy sets. J Am Stat Assoc 99:867–877
32. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall
33. Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9:60–62
34. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
35. Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 23:421–427
36. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8:199–249, 8:301–357, 9:43–80

# Combining the Information of Multiple Ranker in Ranked Set Sampling with Fuzzy Set Approach

**Bekir Cetintav, Selma Gurler, Neslihan Demirel and Gozde Ulutagay**

**Abstract** Ranked set sampling (RSS) is a useful alternative sampling method for parameter estimation. Compared to other sampling methods, it uses the ranking information of the units in the ranking mechanism before the actual measurement. The ranking mechanism can be described as a visual inspection of an expert or a highly-correlated concomitant variable. Accuracy for ranking of the sample units affects the precision of the estimation. This study proposes an alternative approach, called Fuzzy-weighted Ranked Set Sampling (FwRSS), to RSS for dealing with the uncertainty in ranking using fuzzy set. It assumes that there are $K$ $(K > 1)$ rankers for rank decisions and uses three different fuzzy norm operators to combine the decisions of all rankers in order to provide the accuracy of ranking. A simulation study is constructed to see the performance of the mean estimators based on RSS and FwRSS.

## 1 Introduction

In a scientific research, sampling method plays an important role in collecting data set fitting their intended uses in the research. Ranked set sampling (RSS) is an advanced and effective method to obtain data for getting information and inference about the population. The main impact of RSS is to use the ranking information of

B. Cetintav (✉) · S. Gurler · N. Demirel
Faculty of Science, Department of Statistics, Dokuz Eylul University,
Buca, Izmir, Turkey
e-mail: bekir.cetintav@deu.edu.tr

S. Gurler
e-mail: selma.erdogan@deu.edu.tr

N. Demirel
e-mail: neslihan.ortabas@deu.edu.tr

G. Ulutagay
Faculty of Engineering, Department of Industrial Engineering,
Izmir University, Uckuyular, Izmir, Turkey
e-mail: gozde.ulutagay@izmir.edu.tr

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_15

the units in the sampling mechanism. When the ranking is done properly with an expert judgment or concomitant variable, the inference based on RSS generally gives better results comparing with simple random sampling (SRS) for both parametric and non-parametric cases [3, 7]. For detailed information see Chen et al. [2].

In RSS, the ranking process is done without actual measurement. Because, there is an ambiguity in discriminating the rank of one unit with another, ranking the units could not be perfect and it may cause uncertainty. There are some studies focused on the modeling the uncertainty with a probabilistic perspective in the literature, such as Dell and Clutter [3], Bohn and Wolfe [1], Ozturk [5], Ozturk [6]. Zadeh [8] introduced fuzzy sets for representing and manipulating data when they are not precise. In our study, we propose to use the fuzzy set theory in the ranking mechanism of RSS under the idea that the units in the ranked sets could belong to not only the most possibly rank but also the other possible ranks. Since fuzzy sets allow the units to belong to different sets with different membership degrees [4], they can be used for modeling the uncertainty in the ranking mechanism as a good way to reduce the ranking error. Another way of reducing the ranking error is using multiple rankers as in Ozturk [6] and combining the ranking information of $K$ ($K > 1$) rankers with a reasonable way. Therefore this shared wisdom can be used to determine which unit will be sampled in a set.

In this study, we propose a fuzzy set approach for modeling the uncertainty in ranking and for combining the information coming from multiple rankers. The new sampling method, Fuzzy-weighted Ranked Set Sampling (FwRSS), enhances the accuracy of ranking using fuzzy sets for rank decisions of each ranker and using three different fuzzy norm operators to combine the decisions of all rankers. We construct a comparative simulation study to show that our new method provides a considerable amount of improvement on the estimation of the population mean over the counterparts in the literature.

## 2 Fuzzy-Weighted Ranked Set Sampling Procedure

In RSS, ranking of a unit in a specific set is performed with a unit having the highest possible ranking order. However there could be naturally some other possible ranks for that unit because the ranking is done without actual measurement. In the fuzzy set concept, we propose to deal with this uncertain situation using memberships of fuzzy sets. Membership degrees provide a mechanism for measuring the degree of membership as a function which represented by a real number in the range [0, 1]. In FwRSS, the membership degrees for rankers can be introduced in two ways. In the first way, the main role of the membership function is to represent human perceptions and decisions as a member of a fuzzy set. As a second way, when a specific concomitant variable is used for ranking the units in the sets, a membership function based on distance can be used to determine the membership degrees. When the distances between units increase, the ranks of the units are determined more clearly and the accuracy of the decisions about the ranks increases. Therefore we propose a membership function which is determined inversely proportional to the distance between the values of concomitant variables for each sampled unit in the set.

Let $r = 1, 2, \dots, m$ be the rank number, $h = 1, 2, \dots, m$ be the set size and $j = 1, 2, \dots, n$ be the cycle number. Let also $Y^k_{(h)j}$ denote the value of $k$th concomitant variable $Y^k$ for $h$th ordered unit of the set in $j$th cycle where $k = 1, 2, \dots, K$. Given that $X_{[h]j}$ is the value of the $h$th ranked unit which is chosen to the sample from the $j$th cycle. The membership degree of $k$th ranker for fuzzy set of rank $r$ and for $h$th ranked unit in a set of $j$th cycle, which is denoted by $m^k_{h,j}(r)$, can be given as follows.

$$m^k_{h,j}(r) = \begin{cases} 1 & \text{for } h = r \\ 1 - \dfrac{|Y^k_{(r)j} - Y^k_{(h)j}|}{\max\limits_{q} |Y^k_{(q)j} - Y^k_{(h)j}|} & \text{for } h \neq r \text{ and } q = 1, 2, \dots, m \end{cases} \tag{1}$$

Note that the membership degrees of a chosen unit for each rank are decided by taking the other units in the set into account. Now, we can describe the seven-step sampling procedure of FwRSS with two main outputs, which are the units chosen to sample and their membership degrees:

1. Select $m$ units at random from a specified population.
2. Each ranker (expert or concomitant variable) ranks these $m$ units without measuring them and determine the membership degrees of the unit for each rank.
3. Combine the membership degree decisions of the rankers.
4. Chose the unit which has the highest membership degree for first rank and retain its membership degrees.
5. Select another $m$ units at random, ranks these units, determine the membership degrees and chose the unit which has the highest membership degree for second rank and retain its membership degrees.
6. Continue to this process until $m$ ranked units are chosen for $m$ rank. The first six steps are called a cycle.
7. First six steps are repeated for $n$ times to get $n$ cycle and $m * n$ observations.

In Step 3, the ranking information of multiple rankers could be combined in the frame of the fuzzy set theory using one of the set operations as $min$, $max$ and $average$ operators. $min$ operator can be defined as a conservative or pessimistic combiner which takes only minimums, $max$ can be defined as a liberal or optimistic combiner which takes the maximums and $average$ operator can be defined as a balanced operator which takes the averages (for detailed information about fuzzy norm operators, see Zimmermann [9]). In our study, we will use each of these operators to obtain the set of combined membership degree of $h$th ordered unit, say $m_{h,j}$, given as follows.

$$m_{h,j}(r) = \begin{cases} min(m^1_{h,j}(r), \ldots, m^K_{h,j}(r)), & \text{if } min \text{ is chosen} \\ max(m^1_{h,j}(r), \ldots, m^K_{h,j}(r)), & \text{if } max \text{ is chosen} \\ average(m^1_{h,j}(r), \ldots, m^K_{h,j}(r)), & \text{if } average \text{ is chosen} \end{cases} \tag{2}$$

## 3 Estimation of Population Mean

From the FwRSS procedure, we will obtain two outputs. First one is the observations and second one is the membership degree matrix consists of the membership degrees of the observations to the each rank. Let $O_j$ be chosen units from $j$th cycle and $M_j$ be their membership matrix consist of the membership degrees (combined via chosen operator) of the sampled units. Illustration for a specific cycle $j$ is given below for $m = 3$ (Fig. 1).

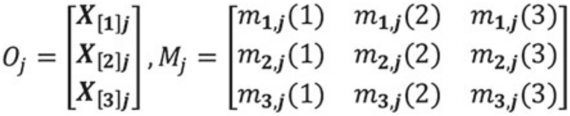
By using the measured $X_{[h]j}$ values of the sampled units and their membership degrees, $m_{(h,j)}(r)$, we define a new estimator for the population mean as:

$$\bar{X}_{FwRSS} = \frac{1}{m} \sum_{r=1}^{m} \sum_{h=1}^{m} \sum_{j=1}^{n} \frac{m_{h,j}(r) X_{[h]j}}{\sum_{h=1}^{m} \sum_{j=1}^{n} m_{h,j}(r)}$$

where $r = 1, 2, \ldots, m$ is the rank number, $h = 1, 2, \ldots, m$ is the set size and $j = 1, 2, \ldots, n$ is the cycle number. Equation given above is a general form of the estimator. In RSS notation, each rank $r = 1, 2, \ldots, m$ are supposed as a stratum and mean of each stratum are estimated for the estimation of population mean. Thus we can rewrite the formula as follows.

$$\bar{X}_{FwRSS} = \frac{1}{m} \sum_{r=1}^{m} \bar{X}^r_{FwRSS} \quad where \quad \bar{X}^r_{FwRSS} = \sum_{h=1}^{m} \sum_{j=1}^{n} \frac{m_{h,j}(r) X_{[h]j}}{\sum_{h=1}^{m} \sum_{j=1}^{n} m_{h,j}(r)}$$

As it is shown in the formula given above, the membership degrees of the units for the ranks $r = 1, 2, \ldots, m$ are used as weights to calculate the mean of each rank.



$$O_j = \begin{bmatrix} X_{[1]j} \\ X_{[2]j} \\ X_{[3]j} \end{bmatrix}, M_j = \begin{bmatrix} m_{1,j}(1) & m_{1,j}(2) & m_{1,j}(3) \\ m_{2,j}(1) & m_{2,j}(2) & m_{2,j}(3) \\ m_{3,j}(1) & m_{3,j}(2) & m_{3,j}(3) \end{bmatrix}$$

**Fig. 1** Outputs of FwRSS procedure for $j$th cycle

**Table 1** Comparison results for FwRSS versus SRS, RSS and multiple RSS

| | Corr. levels | | | Relative efficiencies | | | | |
|---|---|---|---|---|---|---|---|---|
| Comb. method | $Y_1$ | $Y_2$ | $Y_3$ | SRS | RSS($Y_1$) | RSS($Y_2$) | RSS($Y_3$) | Multiple RSS |
| Average | 0.95 | 0.90 | 0.85 | 2.0836 | 1.1774 | 1.2539 | 1.3603 | 1.0569 |
| | 0.55 | 0.75 | 0.95 | 1.9863 | 1.7108 | 1.4081 | 1.1152 | 1.0454 |
| | 0.35 | 0.65 | 0.95 | 1.7564 | 1.6449 | 1.4228 | 1.0092 | 1.0523 |
| | 0.35 | 0.35 | 0.35 | 1.1228 | 1.0477 | 1.0545 | 1.0698 | 1.0130 |
| Min | 0.95 | 0.90 | 0.85 | 1.8101 | 1.0229 | 1.10509 | 1.1807 | 0.9181 |
| | 0.55 | 0.75 | 0.95 | 1.5611 | 1.3446 | 1.1067 | 0.8765 | 0.8216 |
| | 0.35 | 0.65 | 0.95 | 1.4243 | 1.3338 | 1.1538 | 0.8183 | 0.8533 |
| | 0.35 | 0.35 | 0.35 | 0.8936 | 0.8339 | 0.8470 | 0.8514 | 0.8062 |
| Max | 0.95 | 0.90 | 0.85 | 1.9803 | 1.1191 | 1.2089 | 1.2928 | 1.0045 |
| | 0.55 | 0.75 | 0.95 | 1.5382 | 1.3249 | 1.0905 | 0.8636 | 0.8096 |
| | 0.35 | 0.65 | 0.95 | 1.3102 | 1.2270 | 1.0614 | 0.7528 | 0.7850 |
| | 0.35 | 0.35 | 0.35 | 1.0501 | 0.9799 | 0.9953 | 1.0006 | 0.9474 |

## 4 Simulation Study

In order to see the performance of our new method, a simulation study is modeled through the Dell and Clutter [3], which is widely used in RSS studies. There are symmetric and asymmetric distributions generated as the population of the random variables. MATLAB is used for the simulation and reputation number is 10000. The preliminary results of simulation study (in terms of the relative efficiencies) are summarized in Table 1. RSS($Y_1$) means the concomitant variable $Y_1$ is used as a single ranker in classical RSS. Multiple RSS means the combination method and mean estimator given by Ozturk [6] is used.

## 5 Conclusions

In this study a new sampling method, Fuzzy-weighted Ranked Set Sampling (FwRSS), enhances the ranking accuracy using fuzzy sets for rank decisions of individual ranker and using three different fuzzy norm operators to combine the decisions of all rankers. We define a new estimator for the population mean based on FwRSS. A comparative simulation study is constructed to see the performance of our new method. The preliminary results indicate that our average method is more efficient than the SRS, RSS and Multiple RSS methods for estimation of the population mean. However min and max methods are not as efficient as average method even if they are more efficient than SRS and RSS in most cases.

**Acknowledgments** This study is supported by the Scientific and Technological Research Council of Turkey (TUBITAK-COST Grant No. 115F300) under ISCH COST Action IS1304.

## References

1. Bohn LL, Wolfe DA (1994) The effect of imperfect judgment rankings on properties or procedures based on the ranked-set samples analog of the Mann- Whitney-Wilcoxon statistics. J Am Stat Assoc 89:168–176
2. Chen Z, Bai Z, Sinha BK (2004) Ranked set sampling: theory and application. Springer, New York
3. Dell TR, Clutter JL (1972) Ranked set sampling theory with order statistics background. Biometrika 28:545–555
4. Dubois D, Prade H (1978) Operations on fuzzy numbers. Int J Syst Sci 9(6):613–626
5. Ozturk O (2010) Nonparametric maximum-likelihood estimation of within-set ranking errors in ranked set sampling. J Nonparametric Stat 22(7):823–840
6. Ozturk O (2012) Combining ranking information in judgment post stratified and ranked set sampling designs. Environ Ecol Stat 19:73–93
7. Takahasi K, Wakimoto K (1968) On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann Inst Stat Math 20:1–31
8. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
9. Zimmermann HJ (2010) Fuzzy set theory in wiley interdisciplinary reviews. Comput Stat 2:317–332

# A Savage-Like Representation Theorem for Preferences on Multi-acts

**Giulianella Coletti, Davide Petturiti and Barbara Vantaggi**

**Abstract** We deal with a Savage-like decision problem under uncertainty where, for every state of the world, the consequence of each decision (multi-act) is generally uncertain: the decision maker only knows the set of possible alternatives where it can range (multi-consequence). A Choquet expected utility representation theorem for a preference relation on multi-acts is provided, relying on a state-independent cardinal utility function defined on the (finite) set of all alternatives.

## 1 Introduction

The subjective expected utility (SEU) theory due to Savage [15] is the most known model for decisions under uncertainty, having its roots in the von Neumann-Morgenstern theory [20] for decisions under risk.

As is well-known, Savage's model copes with preferences on acts (or decisions) which are represented through a linear functional, depending on a cardinal utility function and a (unique) non-atomic finitely additive probability. This theory rests upon the assumption that every act associates to each state of the world $s \in S$ a unique consequence $x \in X$ available with certainty. Thus, Savage's model is plainly a "one-stage" decision problem [14] in which there is uncertainty only on the states of the world.

G. Coletti · D. Petturiti (✉)
Dip. Matematica e Informatica, University of Perugia, Perugia, Italy
e-mail: davide.petturiti@dmi.unipg.it

G. Coletti
e-mail: giulianella.coletti@unipg.it

B. Vantaggi
Dip. S.B.A.I., "La Sapienza" University of Rome, Rome, Italy
e-mail: barbara.vantaggi@sbai.uniroma1.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_16

Nevertheless, in many decision problems [11, 13], the consequence of an act for a given state $s \in S$ can be uncertain, though belonging to a known set of possible alternatives $Y \subseteq X$. Such decision problems are naturally "two-stage" as they are characterized by two types of uncertainty: the one on the state of the world that will come true and (for a fixed decision and state) the one on the consequence that will be available.

Such decision problems can be modelled through multi-valued acts (multi-acts) [13] on which the decision maker specifies his preferences. A multi-act $f$ can be seen as a random set [12], which, in turn, corresponds to a set of logical constraints between the Boolean algebras $\wp(S)$ and $\wp(X)$.

From a decision-theoretic point of view, comparing multi-acts can be interpreted as comparing different possible (though not certain) "logical situations" existing between $\wp(S)$ and $\wp(X)$. Hence, choosing a particular multi-act translates in accepting the corresponding set of logical constraints between $\wp(S)$ and $\wp(X)$, which determines the way the uncertainty on $\wp(S)$ is transferred to $\wp(X)$.

In this paper we provide an axiom system for a preference relation on multi-acts to be represented through a Choquet expected utility (CEU) with respect to belief functions, relying on a state-independent cardinal utility function defined on the (finite) set of all alternatives. The resulting Savage-like representation theorem is the counterpart in the context of decisions under uncertainty of the model for decisions under risk given in [4, 5]. Even if our decision framework is common to the one in [13], we are looking for an essentially different representation of the preference relation. Indeed, in [13] the author searches for a linear functional where the utility is defined on subsets of $X$ while here the utility function is defined on $X$ in the spirit of Savage.

In the present model, the second of the two decision stages involving uncertainty is implicitly incorporated in the representation via multi-acts and corresponds to a pessimistic transfer of probabilistic uncertainty on (sets of) consequences. Hence, a CEU maximizer decision maker in the present model realizes a form of maxmin expected utility decision rule [8]. Another well-known "two-stage" CEU model present in the literature is the one in [17] where the second stage has the form of an explicit "objective" probability distribution in the sense of Anscombe-Aumann [1].

## 2 Preliminaries

Let $X = \{x_1, \dots, x_n\}$ be a finite set and denote by $\wp(X)$ the power set of $X$. We recall that a *belief function Bel* [6, 18] on $\wp(X)$ is a function such that $Bel(\emptyset) = 0$, $Bel(X) = 1$ and satisfying the $n$-monotonicity property for every $n \geq 2$, i.e., for every $A_1, \dots, A_n \in \mathcal{A}$,

$$Bel\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\dots,n\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right).$$

Previous properties imply the monotonicity of *Bel* with respect to set inclusion $\subseteq$, hence belief functions are particular *normalized capacities* [3].

A belief function *Bel* on $\wp(X)$ is completely singled out by its *Möbius inverse* (see, e.g., [2]) defined for every $A \in \wp(X)$ as

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B).$$

Such a function, usually called *basic (probability) assignment*, is a function $m : \wp(X) \to [0, 1]$ satisfying $m(\emptyset) = 0$ and $\sum_{A \in \wp(X)} m(A) = 1$, and is such that for every $A \in \wp(X)$

$$Bel(A) = \sum_{B \subseteq A} m(B).$$

A set $A$ in $\wp(X)$ is a *focal element* for $m$ (and so also for the corresponding *Bel*) whenever $m(A) > 0$. In particular, a belief function is a probability measure if all its focal elements are singletons.

Recall that every multi-valued mapping from a finitely additive probability space $(S, \wp(S), P)$ to $X$ gives rise to a belief function on $\wp(X)$ [6, 19].

For a function $u : X \to \mathbb{R}$ such that $u(x_1) \leq \ldots \leq u(x_n)$ the *Choquet integral* of $u$ with respect to *Bel* (see [7]) is defined as

$$\oint u \, \mathrm{d}Bel = \sum_{i=1}^{n} u(x_i)(Bel(E_i) - Bel(E_{i+1})),$$

where $E_i = \{x_i, \ldots, x_n\}$ for $i = 1, \ldots, n$, and $E_{n+1} = \emptyset$.

## 3 Preferences on Multi-acts

Consider the following decision-theoretic setting:

- $S$, a set of states of the world;
- $X = \{x_1, \ldots, x_n\}$, a finite set of consequences;
- $\mathcal{Y} = \wp(X) \setminus \{\emptyset\}$, the set of all multi-consequences on $X$.

**Definition 1** A **multi-act** is a function $f : S \to \mathcal{Y}$. Let us denote with $\mathcal{F} = \mathcal{Y}^S$ the set of all multi-acts among which we distinguish, for every $Y \in \mathcal{Y}$, the **constant multi-act** $\overline{Y}$ defined as

$$\overline{Y}(s) = Y, \quad \text{for every } s \in S.$$

For $Y, Z \in \mathcal{Y}$ and $A \in \wp(S)$, a **bivariate multi-act** $f_A^{Y,Z}$ is defined as

$$f_A^{Y,Z}(s) = \begin{cases} Z & \text{if } s \in A, \\ Y & \text{if } s \in A^c. \end{cases}$$

Multi-consequences are also called *opportunity sets* or *menus* in [11], while multi-acts are referred to as *opportunity acts* in [13].

*Example 1* Let us take an urn $U$ from which we draw a ball. $U$ contains white, black and red balls, but in a ratio entirely unknown to us. Let $S = \{s_1, s_2\}$ with $s_1 =$ "the drawn ball is white" and $s_2 =$ "the drawn ball is not white".

Examples of multi-acts where $X = \{\text{bike, car, boat}\}$ are:

| | $s_1$ | $s_2$ |
|---|---|---|
| $f_1$ | {car} | {car, bike} |
| $f_2$ | {car, boat} | {car, bike} |
| $f_3$ | {bike} | {bike} |

■

Given a finitely additive probability $P$ on $\wp(S)$, every $f \in \mathcal{F}$ induces a probability distribution $P \circ f^{-1}$ on $\mathcal{Y}$ which, in turn, gives rise to a belief function $Bel_f$ on $\wp(X)$ defined as

$$Bel_f(\emptyset) = 0 \quad \text{and} \quad Bel_f(Y) = \sum_{\emptyset \neq Z \subseteq Y} P(f^{-1}(Z)), \quad \text{for every } Y \in \mathcal{Y}. \tag{1}$$

Let us consider a preference relation $\precsim$ on $\mathcal{F}$. Following Savage's construction [15], the aim is to find:

- a (unique) non-atomic finitely additive probability $P : \wp(S) \to [0, 1]$ such that the set of multi-acts $\mathcal{F}$ induces the set **B** all belief functions on $\wp(X)$;
- a (unique up to positive linear transformations) utility function $u : X \to \mathbb{R}$ such that, for every $f, g \in \mathcal{F}, f \precsim g \Longleftrightarrow \oint u \, \mathrm{d}Bel_f \leq \oint u \, \mathrm{d}Bel_g$.

Notice that even if the decision framework is common to the one in [13], we are looking for an essentially different representation of the preference $\precsim$ on $\mathcal{F}$. Indeed, in [13] the author searches for a linear functional in the spirit of [9, 10], where the utility is defined on $\mathcal{Y}$ and is a 2-alternating function, the latter having a suitable (non-unique) weighted mean representation as in [11]. On the contrary, here the goal is to find a state-independent cardinal utility function $u$ defined on $X$ in the spirit of Savage.

**Definition 2** For $A \in \wp(S)$, the **conditional preference on** $A$ generated by $\precsim$ is the relation $\precsim_A$ defined, for $f, g \in \mathcal{F}$, as

$$f \precsim_A g \Longleftrightarrow f' \precsim g', \text{ for all } f', g' \in \mathcal{F} \text{ with } f'_{|A} = f_{|A}, g'_{|A} = g_{|A}, f'_{|A^c} = g'_{|A^c}.$$

The previous relation allows to determine null events.

**Definition 3** An event $A \in \wp(S)$ is **null** if and only if $f \precsim_A g$ for every $f, g \in \mathcal{F}$.

Let us consider the following axioms for $\precsim$ on $\mathcal{F}$:

**(B1)** $\precsim$ is a weak order on $\mathcal{F}$.

**(B2)** For all $f, g, f', g' \in \mathcal{F}$, for all $A \in \wp(S)$, if $\left[f_{|A} = f'_{|A}, g_{|A} = g'_{|A}, f_{|A^c} = g_{|A^c}, f'_{|A^c} = g'_{|A^c}\right]$, then $\left[f \precsim g \Longleftrightarrow f' \precsim g'\right]$.

**(B3)** For all $f, g \in \mathcal{F}$, for all $Y, Z \in \mathcal{Y}$, for all not null $A \in \wp(S)$, if $\left[f_{|A} = \overline{Y}_{|A}, g_{|A} = \overline{Z}_{|A}\right]$, then $\left[f \precsim_A g \Longleftrightarrow \overline{Y} \precsim \overline{Z}\right]$.

**(B4)** For all $Y, Z, V, W \in \mathcal{Y}$, for all $A, B \in \wp(S)$, if $\left[\overline{Y} \prec \overline{Z}, \overline{V} \prec \overline{W}\right]$, then $\left[f_A^{Y,Z} \precsim f_B^{Y,Z} \Longleftrightarrow f_A^{V,W} \precsim f_B^{V,W}\right]$.

**(B5)** There are $x_i, x_j \in X$ such that $\overline{\{x_i\}} \prec \overline{\{x_j\}}$.

**(B6)** For all $f, g \in \mathcal{F}$, if $f \prec g$ and $Y \in \mathcal{Y}$, then there is a partition $\{E_1, \ldots, E_m\}$ of $S$ such that for $i = 1, \ldots, m$

$$\begin{cases} \left[f'_{E_i} = \overline{Y}_{|E_i}, f'_{|E_i^c} = f_{|E_i^c}\right] \Longrightarrow f' \prec g, \\ \left[g'_{E_i} = \overline{Y}_{|E_i}, g'_{|E_i^c} = g_{|E_i^c}\right] \Longrightarrow f \prec g'. \end{cases}$$

The following relations are induced by $\precsim$.

**Definition 4** The **preference relation on consequences** $\leq^*$ on $X$ induced by $\precsim$ is defined, for $x_i, x_j \in X$, as

$$x_i \leq^* x_j \Longleftrightarrow \overline{\{x_i\}} \precsim \overline{\{x_j\}}.$$

The **qualitative probability** $\precsim^*$ on $\wp(S)$ induced by $\precsim$ is defined, for $A, B \in \wp(S)$, as

$$A \precsim^* B \Longleftrightarrow f_A^{Y,Z} \precsim f_B^{Y,Z},$$

with $Y, Z \in \mathcal{Y}$ such that $\overline{Y} \prec \overline{Z}$.

If $\precsim$ satisfies (**B1**)–(**B6**) then $\leq^*$ is a weak order with asymmetric part $<^*$ and symmetric part $=^*$, and we can assume $x_1 \leq^* \ldots \leq^* x_n$. Then, we can consider $X^* = X_{/=^*} = \{[x_{i_1}], \ldots, [x_{i_m}]\}$ for which $<^*$ is a strict order, and we can assume $[x_{i_1}] <^* \cdots <^* [x_{i_m}]$.

Axioms (**B1**)–(**B6**) also imply that $\precsim^*$ is a weak order on $\wp(S)$, whose asymmetric and symmetric parts are denoted as $\prec^*$ and $\sim^*$, respectively. Recall that a finitely additive probability $P$ on $\wp(S)$ *represents* $\precsim^*$ if, for every $A, B \in \wp(S), A \precsim^* B \Longleftrightarrow P(A) \leq P(B)$.

The set **B** of all belief functions on $\wp(X)$ contains, in particular, the set **P** of all probability measures on $\wp(X)$. As is well-known [16], the Choquet integral of a function $u : X \to \mathbb{R}$ with respect to an element $Bel_f \in \mathbf{B}$ coincides with the Choquet integral of $u$ with respect to an "extremal" element of **P** dominating $Bel_f$. Thus, in analogy to what happens in [14], the behaviour of a CEU maximizer decision maker in the present model is completely determined by its "additive behaviour" on the set **P**.

For this reason, the following condition (**B7**) can be interpreted as an *extremality axiom* which consistently propagates the "additive behaviour" from **P** to the whole set **B**.

(**B7**) For every $f, g \in \mathcal{F}$, if for $j = 1, \dots, m$ we have

$$\bigcup_{x_i \in [x_{i_j}]} \bigcup_{\{x_i\} \subseteq B \subseteq E_i} f^{-1}(B) \sim^* \bigcup_{x_i \in [x_{i_j}]} \bigcup_{\{x_i\} \subseteq B \subseteq E_i} g^{-1}(B),$$

then $f \sim g$, where $E_i = \{x_i, \dots, x_n\}$ for $i = 1, \dots, n$.

**Theorem 1** *Let $S$ be a set of states of nature, $X = \{x_1, \dots, x_n\}$ a set of consequences, $\mathcal{Y} = \wp(X) \setminus \{\emptyset\}$ the corresponding set of multi-consequences, and $\mathcal{F} = \mathcal{Y}^S$ the set of all multi-acts. If a binary relation $\precsim$ on $\mathcal{F}$ satisfies (**B1**)–(**B7**) then:*

*(i) there is a (unique) non-atomic finitely additive probability $P : \wp(S) \to [0, 1]$ which represents $\precsim^*$ on $\wp(S)$, and $\mathbf{B} = \{Bel_f : \wp(X) \to [0, 1] \,|\, f \in \mathcal{F}\}$, where $Bel_f$ is defined as in (1), corresponds to the set of all belief functions on $\wp(X)$;*
*(ii) there is a (unique up to positive linear transformations) utility function $u : X \to \mathbb{R}$ such that, for every $f, g \in \mathcal{F}$,*

$$f \precsim g \iff \oint u \, \mathrm{d}Bel_f \leq \oint u \, \mathrm{d}Bel_g.$$

*Proof* Every multi-act in $\mathcal{F}$ can be considered as an ordinary Savage act with consequences in $\mathcal{Y}$. Under this interpretation, axioms (**B1**)–(**B4**) and (**B6**) exactly coincide with Savage axioms (**P1**)–(**P4**) and (**P6**), while (**B5**) implies (**P5**) [15]. By Savage representation theorem, there is a (unique) non-atomic finitely additive probability $P : \wp(S) \to [0, 1]$ which represents $\precsim^*$ on $\wp(S)$. This implies that $S$ is uncountable and that $P(\{s\}) = 0$ for every $s \in S$.

The probability measure $P$ induces the set

$$\mathbf{M} = \left\{ m_f : \mathcal{Y} \to [0, 1] \,\middle|\, m_f(Y) = P(f^{-1}(Y)), \, Y \in \mathcal{Y}, f \in \mathcal{F} \right\}$$

of all probability distributions on $\mathcal{Y}$ and, in turn, this implies that the set **B** contains all belief functions on $\wp(X)$. Savage theorem also implies that, for $f, g \in \mathcal{F}$, if $m_f = m_g$ then $f \sim g$. The relation $\precsim$ can be transported to **M** setting $m_f \precsim m_g$ if and only if $f \precsim g$, for $f, g \in \mathcal{F}$.

Let us consider the subset **P** of **M** defined as

$$\mathbf{P} = \{m_f \in \mathbf{M} : \, m_f(Y) = 0, \, \text{card}\, Y > 1, \, Y \in \mathcal{Y}\}$$

which corresponds to the set of all probability distributions on $X$. By Savage representation theorem it follows that **P** is a mixture set and that the restriction of $\precsim$ on **P** satisfies the von Neumann-Morgenstern axioms [20]. So, by the von

Neumann-Morgenstern representation theorem there is a (unique up to positive linear transformations) utility function $u : X \to \mathbb{R}$ such that for every $m_f, m_g \in \mathbf{P}$

$$m_f \precsim m_g \Longleftrightarrow \sum_{i=1}^{n} u(x_i) m_f(\{x_i\}) \leq \sum_{i=1}^{n} u(x_i) m_g(\{x_i\}).$$

Now, consider $X^* = X_{/=^*} = \{[x_{i_1}], \ldots, [x_{i_m}]\}$ and assume without loss of generality $[x_{i_1}] <^* \cdots <^* [x_{i_m}]$. Notice that the utility function $u$ is strictly increasing with respect to the strict preference $<^*$ on $X$, and is constant on each $[x_{i_j}]$, for $j = 1, \ldots, m$.

For an arbitrary multi-act $f \in \mathcal{F}$, we can consider the function $M_f : X^* \to [0, 1]$ defined for every $[x_{i_j}] \in X^*$, as

$$M_f([x_{i_j}]) = \sum_{x_i \in [x_{i_j}]} \sum_{\{x_i\} \subseteq B \subseteq E_i} m_f(B),$$

where $E_i = \{x_i, \ldots, x_n\}$ for $i = 1, \ldots, n$. Note that $M_f([x_{i_j}]) \geq 0$ for every $[x_{i_j}] \in X^*$ and $\sum_{j=1}^{m} M_f([x_{i_j}]) = 1$, thus $M_f$ determines a probability distribution on $X^*$. It is easily proven (see [4, 5]) that

$$\oint u \, \mathrm{d}Bel_f = \sum_{[x_{i_j}] \in X^*} u(x_{i_j}) M_f([x_{i_j}]).$$

Axiom (**B7**) implies that if $m_f, m_g \in \mathbf{M}$ are such that $M_f = M_g$ then $f \sim g$. In particular, if $m_f, m_g \in \mathbf{P}$ are such that $M_f = M_g$ then

$$\sum_{i=1}^{n} u(x_i) m_f(\{x_i\}) = \sum_{i=1}^{n} u(x_i) m_g(\{x_i\}).$$

Hence, introducing the equivalence relation $\equiv_P$ on $\mathbf{M}$ defined, for $m_f, m_g \in \mathbf{M}$, as $m_f \equiv_P m_g$ if and only if $M_f = M_g$, the set $\mathbf{M}_{/\equiv_P}$ can be identified with

$$\mathbf{P}^* = \{M_f : X^* \to [0, 1] \, : \, f \in \mathcal{F}\}$$

which consists of all probability distributions on $X^*$.

The relation $\precsim$ can be transported to $\mathbf{P}^*$ setting for very $M_f, M_g \in \mathbf{P}^*$, $M_f \precsim M_g$ if and only if $f \precsim g$, for $f, g \in \mathcal{F}$. Since for $m_f, m_g \in \mathbf{P}$ it holds

$$\begin{aligned} m_f \precsim m_g &\Longleftrightarrow M_f \precsim M_g \\ &\Longleftrightarrow \sum_{[x_{i_j}] \in X^*} u(x_{i_j}) M_f([x_{i_j}]) \leq \sum_{[x_{i_j}] \in X^*} u(x_{i_j}) M_g([x_{i_j}]), \end{aligned}$$

it follows that for every $f, g \in \mathcal{F}$

$$f \precsim g \Longleftrightarrow \oint u \, \mathrm{d}Bel_f \leq \oint u \, \mathrm{d}Bel_g,$$

and this concludes the proof. □

**Acknowledgments** Work partially supported by INdAM-GNAMPA through the Project 2015 `U2015/000418` and by the Italian MIUR PRIN 2010-11 `2010FP79LR_003`.

## References

1. Anscombe F, Aumann R (1963) A definition of subjective probability. Ann Math Stat 34(1):199–205
2. Chateauneuf A, Jaffray JY (1989) Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. Math Soc Sci 17(3):263–283
3. Choquet G (1954) Theory of capacities. Ann de l'Inst Fourier 5:131–295
4. Coletti G, Petturiti D, Vantaggi B (2015) Decisions under risk and partial knowledge modelling uncertainty and risk aversion. In: Proceedings of ISPTA 2015, pp 77–86
5. Coletti G, Petturiti D, Vantaggi B (2015) Rationality principles for preferences on belief functions. Kybernetika 51(3):486–507
6. Dempster A (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38(2):325–339
7. Denneberg D (1994) Non-additive measure and integral, series B: mathematical and statistical methods, vol 27. Kluwer Academic Publishers, Dordrecht/Boston/London
8. Gilboa I, Schmeidler D (1989) Maxmin expected utility with non-unique prior. J Math Econ 18(2):141–153
9. Jaffray JY (1989) Linear utility theory for belief functions. Oper Res Lett 8(2):107–112
10. Jaffray JY, Wakker P (1993) Decision making with belief functions: compatibility and incompatibility with the sure-thing principle. J Risk Uncertain 7(3):255–271
11. Kreps D (1979) A representation theorem for "Preference for Flexibility". Econometrica 47(3):565–577
12. Molchanov I (2005) Theory of random sets. Probability and its applications. Springer, London Ltd
13. Nehring K (1999) Preference for flexibility in a savage framework. Econometrica 67(1):101–119
14. Sarin R, Wakker P (1992) A simple axiomatization of nonadditive expected utility. Econometrica 60(6):1255–1272
15. Savage L (1972) The foundations of statistics, 2nd edn. Dover
16. Schmeidler D (1986) Integral representation without additivity. Proc Am Math Soc 97(2):255–261
17. Schmeidler D (1989) Subjective probability and expected utility without additivity. Econometrica 57(3):571–587
18. Shafer G (1976) A mathematical theory of evidence. Princeton University Press
19. Shafer G (1979) Allocations of probability. Ann Probab 7(5):827–839
20. von Neumann J, Morgenstern O (1947) Theory of games and economic behavior, 2nd edn. Princeton University Press

# On Some Functional Characterizations of (Fuzzy) Set-Valued Random Elements

**Ana Colubi and Gil Gonzalez-Rodriguez**

**Abstract** One of the most common spaces to model imprecise data through (fuzzy) sets is that of convex and compact (fuzzy) subsets in $\mathbb{R}^p$. The properties of compactness and convexity allow the identification of such elements by means of the so-called support function, through an embedding into a functional space. This embedding satisfies certain valuable properties, however it is not always intuitive. Recently, an alternative functional representation has been considered for the analysis of imprecise data based on the star-shaped sets theory. The alternative representation admits an easier interpretation in terms of 'location' and 'imprecision', as a generalized idea of the concepts of mid-point and spread of an interval. A comparative study of both functional representations is made, with an emphasis on the structures required for a meaningful statistical analysis from the ontic perspective.

## 1 Introduction

The statistical analysis of (fuzzy) set-valued data from the so-called 'ontic' perspective has frequently been developed as a generalization of the statistics for interval data (see, e.g., [1]). From this 'ontic' perspective, (fuzzy) set-valued data are considered as whole entities, in contrast to the epistemic approach, which considers (fuzzy) set-valued data as imprecise measurements of precise data (see, e.g., [2]). Both the arithmetic and metric structure to handle this 'ontic' data is often based on an extension of the Minkowski arithmetic and the distance between either infima and suprema or mid-points and spreads for intervals. In this way, key concepts such as the expected value or the variability, are naturally defined as an extension of the classical notions within the context of (semi-)linear metric spaces.

The generalization of the concept of interval to $\mathbb{R}^p$ keeps the compactness and convexity properties, and this allows the identification of the contour of the convex

A. Colubi (✉) · G. Gonzalez-Rodriguez
Indurot and Department of Statistics, University of Oviedo, 33007 Oviedo, Spain
e-mail: colubi@uniovi.es

G. Gonzalez-Rodriguez
e-mail: gil@uniovi.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_17

and compact sets in $\mathbb{R}^p$ by means of the support function (see, e.g., [6]). The support function is coherent with the Minkowski arithmetic, but sometimes this is not easy to interpret. In [4] the so-called kernel-radial characterization is investigated as an alternative to the support function based on a representation on polar coordinates. This polar representation is established in the context of the star-shaped sets, and is connected with the developments in [3]. It is coherent with alternative arithmetics and distances generalizing the concepts of location and imprecision in an intuitive way, which are of paramount importance in the considered context.

The aim is to show a comparative study of the support function and the kernel-radial representation through some examples. Methodological and practical similarities and differences of both representations for statistical purposes will be highlighted. The rest of the paper is organized as follows. In Sect. 2 both functional representations are formalized and their graphical visualization is shown for some examples. Section 3 is devoted to the comparison of the corresponding statistical frameworks. Section 4 finalizes with some conclusions.

## 2 The Support Function and the Kernel-Radial Characterization

Since the space of fuzzy sets to be considered is a level-wise extension of (convex and compact) sets, the analysis will focus on $\mathcal{K}_c(\mathbb{R}^p) = \{A \subset \mathbb{R}^p \mid A \neq \emptyset$, compact and convex$\}$. For any $A \in \mathcal{K}_c(\mathbb{R}^p)$, the support function of $A$ is defined as $s_A : \mathbb{S}^{p-1} \to \mathbb{R}$ such that $s_A(u) = \sup_{a \in A} \langle a, u \rangle$ for all $u \in \mathbb{S}^{p-1}$, where $\mathbb{S}^{p-1}$ stands for the unit sphere in $\mathbb{R}^p$ and $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^p$. The support function $s_A$ is continuous and square-integrable on $\mathbb{S}^{p-1}$ and characterizes the set $A$ (see, e.g., [6]).

On the other hand, let $\mathcal{K}_S(\mathbb{R}^p)$ be the space of *star-shaped sets* of $\mathbb{R}^p$, i.e., the space of the nonempty compact subsets $A \subset \mathbb{R}^p$ so that there exists $c_A \in A$ such that for all $a \in A$, $\lambda c_A + (1 - \lambda)a \in A$, for all $\lambda \in [0, 1]$, that is, all the points of $A$ are 'visible' from $c_A$ (see, e.g., [6]). The set of points $c_A \in A$ fulfilling the above condition is called *kernel* of $A$, ker$(A)$. Each $c_A \in ker(A)$ is considered a *center* of $A$. Obviously, $\mathcal{K}_S(\mathbb{R}) = \mathcal{K}_c(\mathbb{R})$, but for $p > 1$, $\mathcal{K}_c(\mathbb{R}^p) \subset \mathcal{K}_S(\mathbb{R}^p)$.

A star-shaped set $A$ can be characterized by a center $k_A$ (e.g., the center of gravity of the kernel), and the radial function defined on the unit sphere. The radial function identifies the contour by means of the distance to that center, i.e., by means of the polar coordinates (see, e.g., [6]). Formally, the center of gravity is given by the expected value of the uniform distribution on ker$(A)$, that is, $k_A = \int_{\ker(A)} x d\mu_k$, being $\mu_k$ the normalized Lebesgue measure on ker$(A)$. The *radial function* is defined as the mapping $\rho_A : \mathbb{S}^{p-1} \to \mathbb{R}^+$ such that $\rho_A(u) = \sup\{\lambda \geq 0 : k_A + \lambda u \in A\}$.

The radial function is the inverse of the gauge function, which has been used in [3] in the context of fuzzy star-shaped sets. However, in [3] the gauge function was not used as a basis for the arithmetic and the metric structure of the space, but in

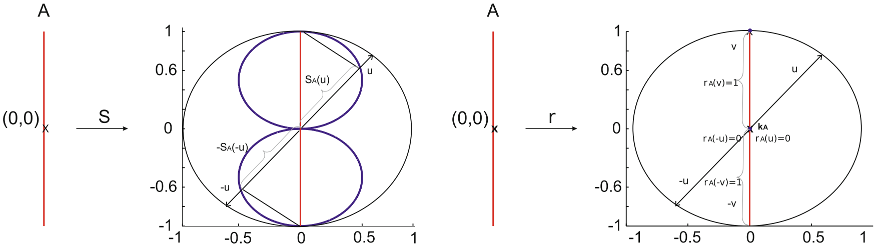


**Fig. 1** Graphical representation of the support function (*left*) and the radial function (*right*) of a line in $\mathcal{K}_c(\mathbb{R}^2)$

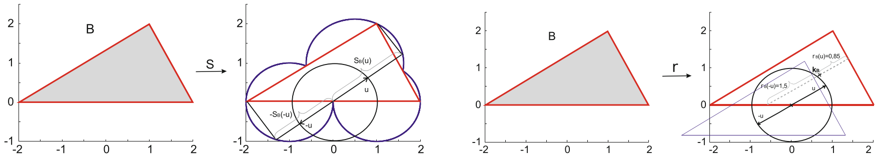


**Fig. 2** Graphical representation of the support function (*left*) and the radial function (*right*) of a *triangle* in $\mathcal{K}_c(\mathbb{R}^2)$

combination with the usual structures, which has reduced the practical usefulness of the proposal.

In order to compare the interpretation of the support and the radial function, Figs. 1 and 2 show a graphical representation of both functions corresponding to a line and a triangle respectively. Since the characterizing functions are defined over the unit sphere, the representations show how each element of the unit sphere relates to the corresponding value. For the support function the sets in $\mathbb{R}^2$ are projected on each one of the directions of the unit sphere and the maximum is computed. In this way, the support function is the distance from the center to the contour of the blue lines. Although this identifies in a unique way the boundaries of the set, the result is not easy to relate with the original shape at first glance. The radial function represents the polar coordinates of the contour line of the original set, that is the radius to each point from the pole (i.e. the steiner point of the kernel). Consequently, the shape of the radial function is straightforwardly connected with the original shape.

For the radial representation, $k_A$ is a center of $A$, describing the location of the set, and $\rho_A$ shows how far the contour line is from this center pointwise. Thus, in line with the idea of mid-point (location) and spread (imprecision) of an interval, $k_A$ and the radial function $\rho_A$ can be identified with the generalized location and imprecision of a star-shaped set respectively.

A previous attempt was made to define generalized concepts of location and imprecision on the basis of the support function by considering the so-called mid-spread representation [7]. This representation is so that $s_A = \text{mid}_A + \text{spr}_A$, where $\text{mid}_A(u) = (s_A(u) - s_A(-u))/2$ and $\text{spr}_A(u) = (s_A(u) + s_A(-u))/2$ for all $u \in \mathbb{S}^{p-1}$. That is, the generalized mid-point/spread is connected with the location/imprecision associated with each direction. This fact entails an interpretational profit, but also some drawbacks from an operational view. Moreover, it inherits the visualization shortcomings from the support function. The main problem is that it is disfficult to determine when a function $s : \mathbb{S}^{p-1} \to \mathbb{R}$ is a support function of any $A \in \mathcal{K}_c(\mathbb{R}^p)$, and this is translated to the mid-function. This problem, however, does not affect the kernel-radial representation, because any function $\rho_A : \mathbb{S}^{p-1} \to \mathbb{R}^+$ is a radial function of a given set.

## 3 Statistical Frameworks

Either through the support function or through the kernel-radial characterization, the space of the corresponding set-valued elements can be embedded into a Hilbert space, namely, $\mathcal{H}_s = L^2(\mathbb{S}^{p-1})$ endowed with the normalized Lebesgue measure on $\mathbb{S}^{p-1}$, $\lambda_p$, for the case of the support function and $\mathcal{H}_r = \mathbb{R}^p \times L^2(\mathbb{S}^{p-1})$ endowed with $\mu_p \times \lambda_p$ for the case of the kernel-radial characterization. Nevertheless, in order to have a meaningful embedding useful for statistical purposes, the arithmetic and metric structures of the original spaces and the Hilbert ones should agree.

It is well known that the support function transfers the Minkowski arithmetic into $\mathcal{H}_s$ and, with the proper metrics, it makes $\mathcal{K}_c(\mathbb{R}^p)$ isometric to a cone of $\mathcal{H}_s$. This arithmetic is defined so that $A +_M \tau B = \{a + \tau b \,|\, a \in A, b \in B\}$ for all $A, B \in \mathcal{K}_c(\mathbb{R}^p)$ and $\tau \in \mathbb{R}$, and verifies that $s_{A+_M\tau B} = s_A + \tau s_B$ for all $\tau \geq 0$. The Minkowski addition is not always meaningful, and there exist various alternatives (see, e.g., [5]).

When the sets are characterized in terms of kernel-radial elements, the natural arithmetic should be coherent as well, that is, $A +_r \tau B$ should be the element in $\mathcal{K}^*(\mathbb{R}^p)$ such that $k^k_{A+_r\tau B} = k_A + \tau k_B$ and $\rho_{A+_r\tau B} = \rho_A + \tau\rho_B$, where the $+$ operator denotes either the usual sum of two points in $\mathbb{R}^p$ or the usual sum of two functions in $L^2(\mathbb{S}^{p-1})$, respectively, for all $A, B \in \mathcal{K}(\mathbb{R}^p)$ and $\tau \in \mathbb{R}$.

Figure 3 shows how sometimes the kernel-radial arithmetic may be more useful than Minkowski's one. The Minkowski and the kernel-radial sum of two lines is shown graphically. The Minkowski sum of two elements in $\mathcal{K}_c(\mathbb{R}^2)$ with null area in $\mathbb{R}^2$ and the same shape results in a convex set with different shape and non-null area. On the contrary, the kernel-radial arithmetic keeps the shape and the surface of the sets.

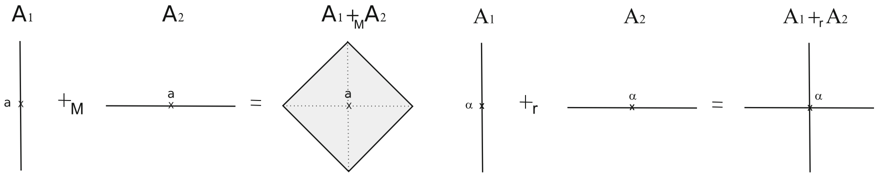


**Fig. 3** Minkowski (*left*) and radial (*right*) sum of two segments

Concerning the metric structure, $L^2$-type metrics are normally considered for statistical purposes. For instance, for the support function-related characterizations, it is common to consider the generalized family for $\theta \in [0, +\infty)$

$$d_\theta(A, B) = \sqrt{||\text{mid}_A - \text{mid}_B||_p^2 + \theta||\text{spr}_A - \text{spr}_B||_p^2},$$

for all $A, B \in \mathcal{K}_c(\mathbb{R}^p)$ where $|| \cdot ||_p$ is the usual $L^2$-type norm for functions defined on $\mathbb{S}^{p-1}$ with respect to $\lambda_p$ [7]. In an analogous way, for the kernel-radial representation, the natural family of metrics for statistical purposes from an ontic point of view is

$$d(A, B) = \sqrt{\tau\|k_A - k_B\|^2 + (1-\tau)\|\rho_A - \rho_B\|_p^2}$$

for all $A, B \in \mathcal{K}(\mathbb{R}^p)$ and $\tau \in (0, 1)$, where $|| \cdot ||$ is the usual Euclidean norm in $\mathbb{R}^p$.

With these structures, it is clear that the considered spaces can be identified with cones of Hilbert spaces, and all the statistical concepts and tools defined in general Hilbert space apply in this context, taking into account that some constraints may arise whenever it is required to remain in the cone. Thus, notions such as random element, expected value, variance or covariance operator, and basic results, such as the CLT, are directly inherited from the theory in Hilbert spaces in the same way for both characterizations. The unique methodological difference in this respect is that, although it is trivial to check if a radial function remains in the cone (i.e. $\rho_A(u) \geq 0$ for all $u \in \mathbb{R}^p$), this is not the case for the support function.

## 4 Conclusions

The support function has traditionally been used to characterize compact and convex sets. This is specially useful when the Minkowski arithmetic is suitable. We have shown that this concept is not always intuitive. As an alternative, the kernel-radial representation is proposed. One of the main advantages of this new representation is that it is easy to interpret in terms of generalized concepts of mid-spread for intervals. The statistical analysis involving both kind of elements can be reduced in both cases to the Hilbert case, so no specific methodology is required to be developed for many common problems. Moreover, the characterization of the cone where the sets are embedded is trivial and similar to the interval case (i.e., non-negativity constraints). This entails a substantial methodological simplification when it is essential to guarantee that any element remains in the cone. Concerning the arithmetic, it has been shown that the Minkowski sum is not always suitable when $p > 1$, as it does not keep shapes or areas, while the arithmetic based on the kernel-radial representation can be a suitable alternative for cases where that is important.

All the discussions in this paper can be extended to the case of fuzzy sets by considering levelwise-defined concepts. Namely, let $\mathcal{F}_c(\mathbb{R}^p)$ be the space of fuzzy sets $U : \mathbb{R}^p \to [0, 1]$ whose $\alpha$-level sets $U_\alpha \in \mathcal{K}_c(\mathbb{R}^p)$ for all $\alpha \in (0, 1]$. Then, the support function can be defined as $s_A : \mathbb{S}^{p-1} \times (0, 1] \to \mathbb{R}$ so that $s_A(u, \alpha) = s_{A_\alpha(u)} \sup_{a \in A_\alpha} \langle a, u \rangle$ for all $u \in \mathbb{S}^{p-1}$, and $\alpha \in (0, 1]$. In the same way, the Minkowski arithmetic is level-wise defined, and the metric is established wrt the joint normalized Lebesgue measure on $\mathbb{S}^{p-1} \times (0, 1]$. Analogous developments can be performed for the case of the kernel-radial representation. The unique technical burden that distinguishes the case of fuzzy sets from the case of standard sets is the problem of building a fuzzy set from the functions on the respective Hilbert spaces, if possible, but this can be done by taking into account the well-known properties that guarantee that a set of indexed levels $\{A_\alpha\}_{\alpha \in [0,1]}$ determines a fuzzy set.

**Acknowledgments** The research in this paper has been partially supported by MTM2013–44212-P, GRUPIN14-005 and the COST Action IC1408. The authors would also like to thank Dr. Angela Blanco for her scientific support.

## References

1. Colubi A, Gonzalez-Rodriguez G (2015) Fuzziness in data analysis: towards accuracy and robustness. Fuzzy Sets Syst 281:260–271
2. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: ontic vs. epistemic views. Int J Approximate Reasoning 55:15021518
3. Diamons P (1990) A note on fuzzy starshaped fuzzy sets. Fuzzy Sets Syst 37:193–199
4. Gonzalez-Rodriguez G, Blanco-Fernandez A, Colubi A (2015) A new framework for the statistical analysis of set- and fuzzy set-valued random elements. Submitted
5. Molchanov I (1998) Averaging of random sets and binary images. Quarterly 11:371–384
6. Schneider R (1993) Convex bodies: the Brunn-Minkowski theory. Cambridge University Press, Cambridge
7. Trutschnig W, Gonzalez-Rodriguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inf. Sci. 179:3964–3972

# Maximum Likelihood Under Incomplete Information: Toward a Comparison of Criteria

**Inés Couso and Didier Dubois**

**Abstract** Maximum likelihood is a standard approach to computing a probability distribution that best fits a given dataset. However, when datasets are incomplete or contain imprecise data, depending on the purpose, a major issue is to properly define the likelihood function to be maximized. This paper compares several proposals in terms of their intuitive appeal, showing their anomalous behavior on examples.

## 1 Introduction

Edwards ([6], p. 9) defines a likelihood function as being proportional to the probability of obtaining results given a hypothesis, according to a probability model. A fundamental axiom is that the probability of obtaining at least one among two results is the sum of the probabilities of obtaining each of these results. In particular, a *result* in the sense of Edwards is not any kind of event, it is an elementary event. Only elementary events can be observed. For instance, when tossing a die, and seeing the outcome, you cannot observe the event "odd", you can only see 1, 3 or 5. If this point of view is accepted, what becomes of the likelihood function under incomplete or imprecise observations? To properly answer this question, one must understand what is a result in this context. Namely, if we are interested in a certain random phenomenon, observations we get in this case do not directly inform us about the underlying random variables. Due to the interference with an imperfect measurement process, observations will be set-valued. So, in order to properly exploit such incomplete information, we must first decide what to model:

1. the random phenomenon *through* its measurement process;
2. or the random phenomenon *despite* its measurement process.

I. Couso (✉)
Department of Statistics, Universidad de Oviedo, Oviedo, Spain
e-mail: couso@uniovi.es

D. Dubois
IRIT, Université Paul Sabatier, Toulouse, France
e-mail: dubois@irit.fr

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_18

In the first case, imprecise observations are considered as results, and we can construct the likelihood function of a random set, whose realizations contain precise but ill-known realizations of the random variable of interest. Actually, most authors are interested in the other point of view, consider that outcomes are the precise, although ill-observed, realizations of the random phenomenon. However in this case there are as many likelihood functions as precise datasets in agreement with the imprecise observations. Authors have proposed several ways of addressing this issue. The most traditional approach is based on the EM algorithm and it comes down to constructing a fake sample of the ill-observed random variable in agreement with the imprecise data, and maximizing likelihood wrt this sample. In this paper we analyze this methodology in the light of the epistemic approach to statistical reasoning outlined in [1] and compare it with several recent proposals by Denoeux [5], Hüllermeier [8], and Guillaume [7]. Note that in this paper we do not consider the issue of imprecision due to too small a number of precise observations (see for instance Serrurier and Prade [10]).

## 2 The Random Phenomenon and Its Measurement Process

Let the random variable $X : \Omega \to \mathcal{X}$ represent the outcome of a certain random experiment. For the sake of simplicity, let us assume that $\mathcal{X} = \{a_1, \ldots, a_m\}$ is finite. Suppose that there is a measurement tool that provides an incomplete report of observations. Namely, the measurement tool reports information items $\Gamma(\omega) = B \in 2^{\mathcal{X}}$, for some multimapping $\Gamma : \Omega \to 2^{\mathcal{X}}$, which represents our (imprecise) perception of $X$, in the sense that we assume that $X$ is a selection of $\Gamma$, i.e. $X(\omega) \in \Gamma(\omega),\ \forall \omega \in \Omega$ [3]. Let $\mathcal{G} = Im(\Gamma) = \{A_1, \ldots, A_r\}$ denote the image of $\Gamma$ (the collection of possible outcomes).

We overview below two different ways to represent the information about the joint distribution of the random vector $(X, \Gamma)$.

**The imprecision generation standpoint**. Here, we emphasize the outcome of the experiment $X$ and the "imprecisiation" process that leads us to just get imprecise observations of $X$, Let us consider the following matrix: $(M|\mathbf{p})$, where $M$ is called the mixing matrix with terms:

- $a_{jk} = p_{.j|k.} = P(\Gamma = A_j|X = a_k)$ denotes the (conditional) probability of observing $A_j$ if the true outcome is $a_k$ and
- $p_{k.} = P(X = a_k)$ denotes the probability that the true outcome is $a_k$.

Such a matrix determines the joint probability distribution modeling the underlying generating process plus the connection between true realizations and incomplete observations. Some examples and their characterizing matrices are as follows:

- **Partition** [4]. Suppose that $Im(\Gamma) = \{A_1, \ldots, A_r\}$ forms a partition of $\mathcal{X}$. Therefore, we can easily observe that the probabilities $P(\Gamma = A_j|X = a_k) = 1$ if $a_k \in A_j$ and 0 otherwise, forall $j, k$.

- **Superset assumption** [9]. $Im(\Gamma)$ coincides with $2^{\mathcal{X}} \setminus \{\emptyset\}$. For each $k = 1, \ldots, m$ there is a constant $c_k$ such that $P(\Gamma = A_j | X = a_k) = c_k$, if $A_j \ni a_k$ $(P(\Gamma = A_j | X = a_k)) = 0$, otherwise.) Furthermore, for every $k \in \{1, \ldots, m\}$ there are $2^{m-1}$ subsets of $\mathcal{X}$ that contain it. Therefore the constant is equal to $1/2^{m-1}$, i.e.: $P(\Gamma = A_j | X = a_k) = \begin{cases} 1/2^{m-1} & ifA_j \ni a_k \\ 0 & otherwise. \end{cases}$ This is a kind of missing-at-random assumption, whereby the imprecisiation process is completely random. It is often presented as capturing the idea of "lack of information" about this process, which sounds questionable.

**The disambiguation standpoint**. We can alternatively characterize the joint probability distribution of $(X, \Gamma)$ by means of the marginal distribution of $\Gamma$ (the mass assignment $m(A_j) = P(\Gamma = A_j) = p_{.j}$, $j = 1, \ldots, r$ of a belief function describing imprecise observations [3]) and the conditional probability of each result $X = a_k$, knowing that the observation was $\Gamma(\omega) = A_j$, for every $j = 1, \ldots, r$. The new matrix $(M' | \mathbf{p}')$ can be written as follows:

- $b_{kj} = p_{k.|.j} = P(X = a_k | \Gamma = A_j |)$ denotes the (conditional) probability that the true value of $X$ is $a_k$ if we have been reported that it belongs to $A_j$
- $p_{.j} = P(Y = A_j) = P(\Gamma = A_j)$ denotes the probability that the generation plus the measurement processes lead us to observe $A_j$.

Such a matrix determines the joint probability distribution modeling the underlying generating process plus the connection between true outcomes and incomplete observations. (More specifically, the vector $(p_{.1}, \ldots, p_{.r})^T$ characterizes the observation process while the matrix $B = (p_{k.|.j})_{k=1,\ldots,m; j=1,\ldots,r}$ represents the conditional probability of $X$ (true outcome) given $\Gamma$ (observation). Here is an example:

- **Uniform conditional distribution** Under the uniform conditional distribution, the (marginal) probability $P_X$ induced by $X$ is the pignistic transform [11] of the belief measure associated to the mass assignment $m$. The conditional distribution is given by: $p_{k.|.j} = \frac{1}{\#A_j}$, if $a_k \in A_j$ and 0 otherwise. And the marginal distribution is: $p_{k.} = \sum_{j: A_j \ni a_k} \frac{1}{\#A_j} p_{.j}$.

## 3 Different Likelihood Functions

Both matrices $M = (A|\mathbf{p})$ and $M' = (B|\mathbf{p}')$ univocally characterize the joint distribution of $(X, \Gamma)$. For each pair $(k, j) \in \{1, \ldots, m\} \times \{1, \ldots, r\}$, let $p_{kj}$ denote the joint probability $p_{kj} = P(X = a_k, \Gamma = A_j)$. According to the nomenclature used in the preceding subsections, the respective marginals on $\mathcal{X}$ and $\mathcal{G}$ are denoted as follows:

- $p_{.j} = \sum_{k=1}^{m} p_{kj}$ will denote the mass of $\Gamma = A_j$, for each $j = 1, \ldots, r$, and
- $p_{k.} = P(X = a_k) = \sum_{j=1}^{r} p_{kj}$ will denote the mass of $X = a_k$, for every $k$.

Now, let us assume that the above joint distribution is characterized by means of a (vector of) parameter(s) $\theta \in \Theta$ (in the sense that $M$ and $M'$ can be written as functions of $\theta$). We naturally assume that the number of components of $\theta$ is less than or equal to the dimension of both matrices, i.e., it is less than or equal to the minimum $\min\{m \times (r+1), r(m+1)\}$. In other words, the approach uses a parametric model such that a value of $\theta$ determines a joint distribution on $\mathcal{X} \times Im(\Gamma)$.

For a sequence of $N$ iid copies of $Z = (X, \Gamma)$, $\mathbf{Z} = ((X_1, \Gamma_1), \ldots, (X_N, \Gamma_N))$, we denote by $\mathbf{z} = ((x_1, G_1), \ldots, (x_N, G_N)) \in (\mathcal{X} \times \mathcal{G})^N$ a specific sample of the vector $(X, \Gamma)$. Thus, $\mathbf{G} = (G_1, \ldots, G_N)$ will denote the observed sample (an observation of the set-valued vector $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_n)$), and $\mathbf{x} = (x_1, \ldots, x_N)$ will denote an arbitrary artificial sample from $\mathcal{X}$ for the unobservable latent variable $X$, that we shall vary in $\mathcal{X}^N$. The samples $\mathbf{x}$ are chosen such that the number of repetitions $n^{kj}$ of each pair $(a_k, A_j) \in \mathcal{X} \times \mathcal{G}$ in the sample are in agreement with the numbers $n_{.j}$ of observations $A_j$. We denote by $\mathcal{X}^{\mathbf{G}}$ (resp. $\mathcal{Z}^{\mathbf{G}}$), the set of samples $\mathbf{x}$ (resp. complete joint samples $\mathbf{z}$) respecting this condition.We may consider three different log-likelihood functions depending on whether we refer to

- the observed sample: $L^{\mathbf{G}}(\theta) = \log \mathbf{p}(\mathbf{G}; \theta) = \log \prod_{i=1}^N p(G_i; \theta)$. It also writes $= \sum_{j=1}^r n_{.j} \log p_{.j}^\theta$. where $n_{.j}$ denotes the number of repetitions of $A_j$ in the sample of size $N$
- the (ill-observed) sample of outcomes: $L^{\mathbf{x}}(\theta) = \log \mathbf{p}(\mathbf{x}, \theta)$. It also writes $\log \prod_{i=1}^N p(x_i; \theta) = \sum_{k=1}^m n_{k.} \log p_{k.}^\theta$, where $n_{k.}$ denotes the number of occurrences of $a_k$ in the sample $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}^{\mathbf{G}}$.
- the complete sample: $L^{\mathbf{z}}(\theta) = \log \mathbf{p}(\mathbf{z}, \theta) = \log \prod_{i=1}^N p(z_i; \theta)$. It also writes $\sum_{k=1}^m \sum_{j=1}^r n_{kj} \log p_{kj}^\theta$ where $n_{kj} = \sum_{i=1}^N 1_{\{(a_k, A_j)\}}(x_i, G_i)$ denotes the number of repetitions of the pair $(a_k, A_j)$ in the sample (i.e., $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$).

In the sequel, we compare some existing strategies of likelihood maximization, based on a sequence of imprecise observations $\mathbf{G} = (G_1, \ldots, G_N) \in \mathcal{G}^N$:

- The standard maximum likelihood estimation (MLE) : it computes the argument of the maximum of $L^{\mathbf{G}}$ considered as a mapping defined on $\Theta$, i.e.: $\hat{\theta} = \arg\max_{\theta \in \Theta} L^{\mathbf{G}}(\theta) = \arg\max_{\theta \in \Theta} \prod_{j=1}^r (p_{.j}^\theta)^{n_{.j}}$. The result is a mass assignment on $2^{\mathcal{X}}$. For instance, the EM algorithm [4] is an iterative technique using a latent variable $X$ to achieve a local maximum of $L^{\mathbf{G}}$.
- The maximax strategy [8]: it aims at finding the pair $(\mathbf{x}^*, \theta^*) \in \mathcal{X}^{\mathbf{G}} \times \Omega$ that maximizes $L^{\mathbf{z}}(\theta)$, i.e.: $(\mathbf{x}^*, \theta^*) = \arg\max_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}, \theta \in \Theta} L^{\mathbf{z}}(\theta)$, i.e., $\arg\max_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}, \theta \in \Theta} \prod_{k=1}^m \prod_{j=1}^r (p_{kj}^\theta)^{n_{kj}}$.
- The maximin strategy [7]: it aims at finding $\theta_* \in \Theta$ that maximizes $L^-(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}} L^{\mathbf{z}}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}} \sum_{k=1}^m \sum_{j=1}^r n_{kj} \log p_{kj}^\theta$. It is a robust approach that also identifies a fake optimal sample $\mathbf{x}_*$.
- The Evidential EM strategy [5]: It assumes that the data set is uncertain and defined by a mass-function over $2^{\mathcal{X}^N}$. Under the particular situation where it has a single focal element $B \subset \mathcal{X}^N$, with mass $m(B) = 1$, the EEM approach considers the following expression as a likelihood function, given such imprecise data (see

Eq. 16 in [5]): $\mathbf{p}(B;\theta) = P((X_1, \ldots, X_N) \in B; \theta)$. The Evidential EM algorithm is viewed as a variation of the classical EM algorithm in order to select a value of $\theta$ that maximizes the "likelihood" $\mathbf{p}(B;\theta)$. In particular, if we assume that $B$ is a Cartesian product of the sets in the collection $\{A_1, \ldots, A_r\}$ the criterion can be alternatively written as follows: $\mathbf{p}(B;\theta) = \prod_{j=1}^{m} P_\theta(X \in A_j)^{n_{\cdot j}}$. The EEM procedure may not coincide with a maximum likelihood estimation since this criterion is not always in the spirit of a likelihood function, as seen later on. The EM algorithm uses it when the imprecise data forms a partition.

Under some particular conditions about the matrices $M$ and $M'$, some of the above likelihood maximization procedures may coincide or not. In the rest of the paper we provide some examples, focusing on the optimal samples $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$ or $\mathbf{x} \in \mathcal{X}^{\mathbf{G}}$ computed by the methods and that are supposed to disambiguate the imprecise data. Indeed most existing techniques end up with computing a probability distribution on $\mathcal{X}$ or a fake sample achieving an imputation of $X$.

## 4 A Comparison of Estimation-Disambiguation Methods

Let us to compare the potentials and limitations of these approaches. Here we just give a few hints by means of examples.

**EM-based approaches**. Let $\mathcal{P}^{\mathcal{X}^N}$ be the set of all probability measures $P$ we can define on the measurable space $(\mathcal{X}^N, \wp(\mathcal{X}^N))$. The EM algorithm [4] tries to maximize the function $F : \mathcal{P}^{\mathcal{X}^N} \times \Theta \to \mathbb{R}$: $F(\mathbf{P}, \theta) = L^{\mathbf{G}}(\theta) - D(\mathbf{P}, \theta)$, $\forall P \in \mathcal{P}^{\mathcal{X}^N}$, $\theta \in \Theta$, where $\mathbf{p}(\mathbf{x}|\mathbf{G};\theta) = \frac{\mathbf{p}(\mathbf{x}, \mathbf{G};\theta)}{\mathbf{p}(\mathbf{G};\theta)}$, whenever $\mathbf{p}(\mathbf{G};\theta) > 0$. Moreover, $D(\mathbf{P}, \mathbf{P}')$ is the Kullback-Leibler divergence from $\mathbf{P}'$ to $\mathbf{P}$, $\sum_{\mathbf{x} \in \mathcal{X}^N} \mathbf{p}(\mathbf{x}) \log[\frac{\mathbf{p}(\mathbf{x})}{\mathbf{p}'(\mathbf{x})}]$, where $\mathbf{p}$ is the mass function associated to $\mathbf{P}$. It is then clear that $L^{\mathbf{G}}(\theta) \geq F(\mathbf{P}, \theta)$ and that if $\mathbf{P} = \mathbf{P}(\cdot|\mathbf{G};\theta)$, then $F(\mathbf{P}, \theta) = L^{\mathbf{G}}(\theta)$. Given a value $\theta^{(n-1)}$ obtained at the $n-1$ M-step, the E-step actually computes $\mathbf{P}(\cdot|\mathbf{G};\theta^{(n-1)})$ (which is basically like determining a fake sample $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$), and the next M step finds a value of $\theta$ that maximizes $F(\mathbf{P}(\cdot|\mathbf{G};\theta^{(n-1)}), \theta)$, i.e. $L^{\mathbf{G}}(\theta)$ based on the fake sample $\mathbf{z}$. In fact, the EM algorithm iteratively finds a parametric probability model $\mathbf{P}^\theta$ and a probability distribution $\mathbf{P}(\cdot|\mathbf{G};\theta)$ on $\mathcal{X}$, that is in agreement with the data $\mathbf{G}$, such that the divergence from $\mathbf{P}^\theta$ to $\mathbf{P}(\cdot|\mathbf{G};\theta)$ is minimal [2]. $\mathbf{P}^\theta$ is an MLE for the fake sample $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$ in agreement with $\mathbf{P}(\cdot|\mathbf{G};\theta)$, which yields the best imputation of $X$ in this sense. There are situations where the result of the EM algorithm will be questionable [2].

*Example 1* Suppose that a dice is tossed and let $X$ pertain to the result of the trial. The probability distribution of $X$ is a vector $(p_1, \ldots, p_6) \in [0, 1]^6$, with $\sum_{i=1}^{6} p_i = 1$. Suppose after each trial we are told either that the result has been less than or equal to 3 ($A_1$) or greater than or equal to 3 ($A_2$). After each toss, when the actual result ($X$) is 3, the reporter needs to decide $A_1$ or $A_2$. Assume the conditional probability $P(G_n = A_1|X_n = 3)$ is a fixed number $\alpha \in [0, 1]$ for every trial, $n = 1, \ldots, N$. Suppose that we toss the dice $N = 1000$ times and the report tells us $n_{.1} = 300$ times that the result was less than or equal to 3. Let $\theta$ denote the vector $(p_1, p_2, p_3, p_4, p_5; \alpha)$. The likelihood function based on the observed sample $\mathbf{G}$ can be written as: $L^{\mathbf{G}}(\theta) = (p_1 + p_2 + \alpha p_3)^{300} \cdot [1 - (p_1 + p_2 + \alpha p_3)]^{700}$. Such a function is maximized for any vector $\theta$ satisfying the constraint $p_1 + p_2 + \alpha p_3 = 0.3$. If we use the EM algorithm, we get a vector $\theta$ satisfying the above constraints after the first iteration of the M algorithm. We will get a different vector $\theta^{(1)}$, depending on the initial point $\theta^{(0)}$. If we start from $\theta^{(0)} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}; \frac{1}{2})$, we get $\theta^{(1)} = (0.12, 0.12, 0.16, 0.2, 0.2; \frac{3}{8})$. It is also the MLE of $\theta$ based on a (fake) sample of 1000 tosses of the dice where the number of repetitions of each of the six facets has been respectively 120, 120, 160, 200, 200, 200. But this is not the only MLE based on the observed sample.

**Evidential EM Algorithm**. We can distinguish the following cases:

- The case where $Im(\Gamma)$ forms a partition of $\mathcal{X}$. In this case, $P(X \in A_j) = P(Y = A_j) = p_{.j},\ \forall j = 1, \dots, r$, and therefore $\prod_{j=1}^{r} P(X \in A_j; \theta)^{n_{.j}}$ coincides with the likelihood $\mathbf{p}(\mathbf{G}; \theta)$.
- The case where the sets $A_1, \dots, A_r$ do overlap. In this case, $\mathbf{p}(\mathbf{G}; \theta)$ and $\mathbf{p}(B; \theta)$ do not necessarily coincide, as shown in the following example.

*Example 2* Let us take a sample of $N$ tosses of the dice in Example 1 and let us assume that the reporter has told us that $n_1$ of the times the result was less than or equal to 3, and $n_2 = N - n_1$ otherwise. The EEM likelihood is $\mathbf{p}(B; \theta) = (p_1 + p_2 + p_3)^{n_1} \cdot (p_3 + p_4 + p_5 + p_6)^{n_2}$ with $\sum_{i=1}^{6} p_i = 1$. We can easily observe that it reaches its maximum ($\mathbf{p}(B; \theta) = 1$) for any vector $\theta$ satisfying the constraint $p_3 = 1$. But such a prediction of $\theta$ would not be a reasonable estimate for $\theta$.

**The maximax approach**. The parametric estimation based on the maximax approach does not coincide in general with the MLE. Furthermore, it may lead to questionable imputations of $X$.

*Example 3* Let us suppose that a dice is tossed $N = 10$ times, and that Peter reports 4 heads, 2 tails and he does not tell whether there was heads or tails for the remaining 4 times. Let us consider the parameter $\theta = (p, \alpha, \beta)$, where $p = P(X = h)$, $\alpha = P(\Gamma = \{h, t\} | X = h)$ and $\beta = P(\Gamma = \{h, t\} | X = t)$. It determines the following joint probability distribution induced by $(X, \Gamma)$: $P(h, \{h\}) = (1 - \alpha)p$; $P(h, \{h, t\}) = \alpha p$; $P(t, \{t\}) = (1 - \beta)p$; $P(t, \{h, t\}) = \beta p$; and 0 otherwise.

The MLE of $\theta$ is not unique. It corresponds to all the vectors $\theta = (p, \alpha, \beta) \in [0, 1]^3$ satisfying the constraints: $(1 - \alpha)p = 0.4$ and $(1 - \beta)(1 - p) = 0.2$, indicating the marginal probabilities $P(\Gamma = \{h\})$ and $P(\Gamma = \{t\})$ respectively.

In contrast, the maximax strategy seeks for a pair $(\theta^*; \mathbf{x}^*) = (p^*, \alpha^*, \beta^*; \mathbf{x}^*)$ that maximizes $L^{\mathbf{z}}(\theta)$. It can be checked that the tuple that maximizes $L^{\mathbf{z}}(\theta)$ is unique. It corresponds to the vector of parameters $\theta^* = (p^*, \alpha^*, \beta^*) = (0.8, 0.5, 0)$ and the sample where all the unknown outcomes are heads. In words, the maximax strategy assumes that all the ill-observed results correspond to the most frequent observed outcome ("heads"). Accordingly, the estimation of the probability of heads is the corresponding frequency (0.8). According to this strategy, and without having any insight about the behaviour of Peter, we predict that each time he refused to report, the result was in fact "heads".

*Example 4* Let us now consider the situation about the coin described in Example 3, and let us suppose in addition that the following conditions hold: $\alpha = 1 - \alpha = 0.5$ and $\beta = 1 - \beta = 0.5$. In words, no matter what the true outcome is (heads or tails) Peter refuses to give any information about it with probability 0.5 (the behavior of Peter does not depend on the true outcome). This is the "superset assumption" [8] already mentioned. Under this additional constraint, the MLE of $\theta = (p, 0.5, 0.5)$ is reached at $\hat{p} = 4/6 = 2/3$. The maximum likelihood estimator provides the same estimation as if we had just tossed the coin six times, since, as a consequence of the superset assumption here, the four remaining tosses play no role in the statistics. As a result, the conditional probability $P(X = h|\Gamma = \{h, t\})$ is assumed to coincide with $P(X = h|\Gamma \neq \{h, t\})$ and with $P(X = h) = p$. Such a probability is estimated from the six observed outcomes, where four of them were "heads" and the rest were "tails". In contrast, the maximax strategy without the superset assumption leads us to take into account the unobserved tosses as matching the most frequent observed outcome, hence the imputation of $X$ is compatible with a data set containing 8 heads and only 2 tails.

**The maximin approach**. Consider again Example 3. The maximin approach consists of considering all log-likelihood functions $L_k^{\mathbf{x}}(p) = (4 + k)\log p + (6 - k)\log(1 - p)$ with $0 \leq k \leq 4$. The approach consists in finding for each value of $p$ the complete data that minimizes $L^{\mathbf{x}}(p)$. Since $L_k^{\mathbf{x}}(p)$ is of the form $k \log \frac{p}{(1-p)} + a$, it is easy to see that if $p < 1/2$, the minimum $L^{-}(p)$ is reached for $k = 4$, and if $p > 1/2$, it is reached for $k = 0$. So, it is $8\log p + 2\log(1 - p)$ if $p < 1/2$ and $4\log p + 6\log(1 - p)$ otherwise. So $L^{-}(p)$ is increasing when $p < 1/2$ and decreasing when $p > 1/2$. It reaches its maximum for $p = 1/2$. So the maximin approach is cautious in the sense of maximizing entropy in the coin-tossing experiment. It yields the uniform distribution, i.e., an imputation of 5 heads and 5 tails, in agreement with the observations.

## 5 Conclusion

This paper suggests that it is not trivial to extend MLE methods to incomplete data despite the existence of several proposals. In particular, it is very questionable to reconstruct distributions for unobserved variables when parameters of distributions that generate them are not closely connected to parameters of distributions that govern observed ones. In contrast, the famous EM article [4] deals with imprecise observations forming a partition and starts with an example in which a single parameter determines the joint distribution of $X$ and $\Gamma$. However, it is not straightforward to adapt the EM procedure to incomplete overlapping data. In the general case, either one applies standard MLE to observed imprecise data only (yielding a mass function) or one has to add an assumption that comes down to selecting a single probability measure in the credal set induced by this mass function. Each approach to imprecise data MLE proposes its own assumption. As can be seen from the examples, it is easy to find cases where these methods lead to debatable solutions: the solution to the EM algorithm [4] depends on the initial parameter value, the EEM approach [5] seems to optimize a criterion that sometimes does not qualify as a genuine likelihood function, the maximax approach [8] may select a very unbalanced distribution for the hidden variable, while the maximin robust MLE [7] favors uninformative distributions. More work is needed to characterize classes of problems where one estimation method is justified and the other method fails.

**Acknowledgments** This work is partially supported by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02, by TIN2014-56967-R (Spanish Ministry of Science and Innovation) and FC-15-GRUPIN14-073 (Regional Ministry of the Principality of Asturias).

## References

1. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: Ontic vs. epistemic views. Int J Approximate Reasoning 55(7):1502–1518
2. Couso I, Dubois D (2016) Belief revision and the EM algorithm. Proc, IPMU
3. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–339
4. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc B 39:1–38
5. Denoeux T (2013) Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans Knowl Data Eng 26:119–130
6. Edwards AWF (1972) Likelihood. Cambridge University Press
7. Guillaume R, Dubois D (2015) Robust parameter estimation of density functions under fuzzy interval observations, 9th ISIPTA Symposium. Pescara, Italy
8. Hüllermeier E (2014) Learning from imprecise and fuzzy observations. Int J Approximate Reasoning 55(7):1519–1534
9. Hüllermeier E, Cheng W (2015) Superset learning based on generalized loss minimization. ECML/PKDD 2:260–275
10. Serrurier M, Prade H (2013) An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. Int J Approximate Reasoning 54(7):919–933
11. Smets P (2005) Decision making in the TBM: the necessity of the pignistic transformation. Int J Approximate Reasoning 38:133–147

# The Use of Uncertainty to Choose Matching Variables in Statistical Matching

**Marcello D'Orazio, Marco Di Zio and Mauro Scanu**

**Abstract** Statistical matching aims at combining information available in distinct sample surveys referred to the same target population. The matching is usually based on a set of common variables shared by the available data sources. For matching purposes just a subset of all the common variables should be used, the so called matching variables. The paper presents a novel method for selecting the matching variables based on the analysis of the uncertainty characterizing the matching framework. The uncertainty is caused by unavailability of data for estimating parameters describing the association/correlation between variables not jointly observed in a single data source. The paper focuses on the case of categorical variables and presents a sequential procedure for identifying the most effective subset of common variables in reducing the overall uncertainty.

## 1 Introduction

*Statistical matching* (sometimes called *data fusion* or *synthetical matching*) aims at combining information available in distinct sample surveys referred to the same target population. Formally, let $Y$ and $Z$ be two random variables; statistical matching techniques can be applied for estimating the joint $(Y, Z)$ distribution function (e.g., a contingency table or a regression coefficient) or some of its parameters when: (i) $Y$ and $Z$ are not jointly observed in a survey, but $Y$ is observed in a sample $A$, of size $n_A$, and $Z$ is observed in a sample $B$, of size $n_B$; (ii) $A$ and $B$ are independent and units in the two samples do not overlap (it is not possible to use record linkage); (iii) $A$ and $B$ both observe a set of additional variables $X$.

M. D'Orazio · M. Di Zio (✉) · M. Scanu
Istat, via Cesare Balbo 16, 00184 Roma, Italy
e-mail: dizio@istat.it

M. D'Orazio
e-mail: madorazi@istat.it

M. Scanu
e-mail: scanu@istat.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_19

## 2 Choice of the Matching Variables

In statistical matching (SM) the data sources $A$ and $B$ may share many common variables $X$. This is the case of matching of data from household surveys where a very high number of variables concerning the household (living place, housing, number of members, etc.) and its members (age, gender, educational level, professional status, etc.) are available. In performing SM, not all the $X$ variables will be used but just the most important ones. The selection of the most relevant $X_M$ ($X_M \subseteq X$), usually called *matching variables*, should be performed by consulting subject matter experts and through appropriate statistical methods.

The choice of the matching variables should be made in a multivariate sense [4] to identify the subset $X_M$ connected, at the same time, with $Y$ and $Z$. This would require the availability of a data source in which $(X, Y, Z)$ are observed. In the basic SM framework, $A$ permits to investigate the relationship between $Y$ and $X$, while the relationship between $Z$ and $X$ can be studied in $B$. The results of the two separate analyses are then joined and, in general, the following rule can be applied:

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z$$

where $X_Y(X_Y \subseteq X)$ and $X_Z(X_Z \subseteq X)$ are the subsets of the common variables that better explain $Y$ and $Z$, respectively. The intersection $X_Y \cap X_Z$ provides a smaller subset of matching variables if compared to $X_Y \cup X_Z$; this is an important feature in achieving parsimony. For instance, too many matching variables in a distance hot deck SM micro application can introduce undesired additional noise in the final results. Unfortunately, the risk with $X_Y \cap X_Z$ is that most of the predictors of one target variable will be excluded if they are not in the subset of the predictors of the other target variable. For this reason, the final subset of the matching variables $X_M$ is usually a compromise and the contribution of subject matter experts and data analysts is important in order to achieve the "best" subset. Our proposal is to perform a unique analysis for choosing the matching variables by searching the set of common variables that are the most effective in reducing the *uncertainty* between $Y$ and $Z$.

### *2.1 Uncertainty in Statistical Matching*

Due to the nature of the SM problem (i.e., $Y$ and $Z$ are never jointly observed) there is an intrinsic uncertainty: there cannot be unique estimates for the parameters describing the association/correlation between $Y$ and $Z$. Approaches, such as maximum likelihood estimation, offer a set of solutions, all with the same (maximum) likelihood, usually closed, known as likelihood ridge. The non-uniqueness of the solution of the SM problem has been described in different articles (see Chap. 4 in [6] and references therein). Given that $A$ and $B$ do not contain any information on $Y$ and $Z$, apart from their association/correlation with the common variables $X$, the set of solutions describes all the values of the parameters represented by all the possible relationships between $Y$ and $Z$ given the observed data. For this reason, [6] called this set of equally plausible estimates as "the uncertainty set". In order to reduce the uncertainty set, it is necessary to add external information (e.g., a structural zero on a cell of the contingency table of $Y \times Z$ or $Y \times Z|X$ reduce the set of possible values). When $X$, $Y$ and $Z$ are categorical, the uncertainty set can be computed by resorting to the Frèchet bounds. Let $p_{hjk} = Pr(X = h, Y = j, Z = k)$ for $h = 1, \dots, H,\ j = 1, \dots, J,\ k = 1, \dots, K$; by conditioning on $X$, the probability $p_{.jk} = Pr(Y = j, Z = k)$ is in the interval:

$$[\underline{p}_{.jk}, \overline{p}_{.jk}] = \left[\sum_h p_{h..} \max\{0, p_{j|h} + p_{k|h} - 1\}, \sum_h p_{h..} \min\{p_{j|h}, p_{k|h}\}\right] \tag{1}$$

It should be noted that when external information is available, for instance in terms of structural zeros on cells of the contingency table $Y \times Z$) or $Y \times Z|X$ these bounds are not sharp, in fact the admissible values are a closed sub-interval (when the estimated probabilities are compatible), see [9].

The expression (1) allows to derive bounds for each cell in the contingency table $Y \times Z$. This information can be used to derive an overall measure of uncertainty; a very basic one can be obtained by considering the average of the bounds width:

$$d = \frac{1}{J \times K} \sum_{j,k} \left(\overline{p}_{\cdot jk} - \underline{p}_{\cdot jk}\right) \tag{2}$$

This is a simple and straightforward way of measuring uncertainty but it is not unique, for instance alternative measures are proposed in [5].

## 2.2 Choosing the Matching Variables by Uncertainty Reduction

The method proposed for selecting the matching variables when dealing with categorical $X$, $Y$ and $Z$ variables is based on an simple idea: select as matching variables just the subset of the $X$ that are more effective in reducing the uncertainty, measured in our case, in terms of $d$. Unfortunately, the value of $d$ decreases by increasing the number of $X$ variables, even when these variables are slightly associated with one or both of the target variables $Y$ and $Z$; for this reason it is necessary to identify a criterion to decide when to stop introducing additional common variable $X$ among the matching variables, having in mind the parsimony principle. All these considerations lead to the definition of the following sequential procedure.

- **Step (0)** Initial ordering of the $X$ variables according to their ability in minimizing $\hat{d}$, and select the variable with the smallest $\hat{d}$ as matching variable;

- **Step (1)** Consider all the possible combinations obtained by adding one more variable to the selected set of variables and evaluate their uncertainty in terms of $\hat{d}$; e.g., in the first iteration all the possible combinations of the variables identified in step (0) with the remaining ones will be considered.
- **Step (2)** Select the combination of variables which determine the higher decrease of the uncertainty ($\hat{d}$) and go back to step (1).

The procedure ends when the starting tables for estimating the interval (1) become too sparse. Sparseness here is measured in terms of average cell counts; in particular, given that there are two starting data-sets, $A$ and $B$, sparseness is measured by the minimum value of the averages of the cell counts:

$$\bar{n} = \min\left[\frac{n_A}{c_{X_D Y}}, \frac{n_B}{c_{X_D Z}}\right] \tag{3}$$

where $c_{X_D Y}$ and $c_{X_D Z}$ denote the number of cells in the table $X_D \times Y$ and $X_D \times Z$ respectively; and $X_D$ is the variable obtained by cross-classifying the selected $X$ variables ($X_D = X_1 \times X_2 \times \ldots \times X_M$). The rationale is that of considering the subset of matching variables able to "keep the average number of observations from becoming too small" ([3], p. 140). Too small in our case is meant maintaining $\bar{n} > 1$, in other terms the procedure stops when $\bar{n} \leq 1$. Such a stopping criterion is a subjective choice which reflects the broad definition of sparseness in [2]: "contingency tables are said to be sparse when the ratio of the sample size to the number of cells is relatively small". In the proposed procedure, the higher is the number of $X$ variables being considered, the larger are the tables $X_D \times Y$ and $X_D \times Z$ and, consequently, the higher is the risk of having zero counts (empty cells), i.e., sparse tables. Empty cells can be caused by structural zeros (i.e., events that cannot occur) or, more frequently, by sampling zeros (i.e., no observations of an event that can occur). In our procedure, it is not easy to separate structural from sampling zeros and the presence of many empty cells is unappealing when estimating (1), for this reason the procedure stops when the starting tables become too sparse. In literature, many alternative methods have been proposed to measure sparseness, e.g., measures based on the percentage of expected cell frequencies smaller than 1, 5 or 10, or the percentage of observed zero frequencies. Further studies will be devoted to analyse the performance of other sparseness indicators.

It is worth noting that sparseness can be tackled in a different manner by estimating the probabilities of the contingency table with alternative methods. A widespread practice for compensating for empty cells consists in adding a constant to all the cells (frequently used constants are $1/c$, 0.5 or $\sqrt{n}/c$, being $c$ the number of cells in the table). Such procedure has an unpleasant feature, in fact, adding a constant smooths toward independence. An alternative approach, consists in collapsing adjacent categories for one or more variables; unfortunately it requires arbitrary choices and may not solve the problem; in addition the risk is that of decreasing the degree of association between variables. A viable method can be that of applying a pseudo-Bayes estimator of cells' probabilities ([3], Sect. 12) that combines the sample proportions with the model based estimators, being consistent even when the model does not hold [1]. In practice, the relative frequency of the generic cell $h$ is estimated by a weighted average, thus introducing a sort of smoothing of the data. This method is not analysed in this paper, further studies will be dedicated to the evaluation of this alternative approach.

### *2.3 Estimating Cell Bounds*

In the usual SM setting, estimating the bounds for cells in the contingency table $Y \times Z$, requires an estimation of the probabilities:

$$p_{h\cdot\cdot}, \quad p_{j|h}, \quad p_{k|h}; \;\; h = 1, \ldots, H; \; j = 1, \ldots, J; k = 1, \ldots, K. \tag{4}$$

When $A$ and $B$ are simple random samples, (4) are estimated by considering the corresponding sample proportions ([6], p. 24):

$$\hat{p}_h = \frac{n^A_{h..} + n^B_{h..}}{n_A + n_B}, \;\; \hat{p}_{j|h} = \frac{n^A_{hj.}}{n^A_{h..}}, \;\; \hat{p}_{k|h} = \frac{n^B_{h.k}}{n^B_{h..}} \tag{5}$$

where $n^A_{hj.}$ and $n^B_{h.k}$ are the observed marginal tables from $A$ and $B$ respectively, for $h = 1, \ldots, H; \; j = 1, \ldots, J; k = 1, \ldots, K.$

On the contrary, when dealing with data form complex sample surveys involving stratification and/or clustering, (4) have to be estimated by considering the following expressions ([8], Sect. 13.5):

$$\hat{p}_h = \frac{n_A \hat{N}^A_{h..}/\hat{N}_A + n_B \hat{N}^B_{h..}/\hat{N}_B}{n_A + n_B}, \qquad \hat{p}_{j|h} = \frac{\hat{N}^A_{hj.}}{\hat{N}^A_{h..}}, \qquad \hat{p}_{k|h} = \frac{\hat{N}^B_{h.k}}{\hat{N}^B_{h..}} \tag{6}$$

with $h = 1, \ldots, H; \; j = 1, \ldots, J; k = 1, \ldots, K;$ where the generic population count $N^S_t$ is estimated by:

$$\hat{N}^S_t = \sum_{i=1}^{n_S} w^S_i I(u_i = t) \tag{7}$$

where $w^S_i$ is the survey weight assigned to the $i$th unit of sample $S$ (usually reflecting inclusion probabilities in the sample corrected to compensate for nonresponse, coverage errors, etc.), while $I\,() = 1$ if the condition within parentheses is satisfied and 0 otherwise. Sometimes, for practical purposes, the survey weights are rescaled to sum up to the sample size, i.e. $\hat{N}^S = \sum_{i=1}^{n_S} w^S_i = n_S$; this rescaling has no effect on the estimates of the relative frequencies (6). In any case, in complex samples surveys the practice of estimating cells relative frequencies by discarding the survey weights should be avoided because it may provide inaccurate results.

## 3 Application and Results

In the following sub-sections two applications are presented, both cases refer to artificial data.

**Case 1** Bayesian networks are used to generate two artificial samples of size $n_A = n_B = 5000$ sharing 3 binary $X$. The complete Bayesian network for $Y$, $Z$, $X_1$, $X_2$, and $X_3$ in Fig. 1 shows that there is a direct relationship between $Y$ and $Z$, and that they are dependent on $X_1$, $X_2$, and $X_3$, in fact we notice that $Z$ and $X_3$ are marginally independent, but they are dependent conditionally on $X_2$. The latter two networks in Fig. 1 show the marginal Bayesian networks (once $Z$ and $Y$ are marginalized for the pictures in the centre and the left, respectively) and denote that $Y$ $(Z)$ depends directly only on $X_1(X_2)$ when $Z$ $(Y)$ is missing. The intensity of the association of the variables measured through the Cramer's V is medium (around 0.6) between $(X_1, X_2)$, and $(Y, Z)$, while is quite weak between the other variables (around 0.1).

Application of the step (0) of the procedure in Sect. 2.2 suggests that $X_1$ should be considered as the first one ($\hat{d} = 0.1703$), then $X_3(\hat{d} = 0.1911)$ and finally $X_2$ $(\hat{d} = 0.2012)$.

The procedure for selecting the matching variables selects $X_1$, $X_2$, $X_3$ as matching variables (see Table 1), however $X_1$ alone is able to achieve quite the best score in terms of average width of the uncertainty bounds; adding $X_2$ does not improve the result and just a negligible decrease of $\hat{d}$ is achieved by considering all the $X$ variables.

**Case 2**. As a toy example we refer to two artificial samples, $n_A = 3009$ and $n_B = 6686$, generated from the EU-SILC data (data available in [7]). The ordering of the 7 common variables obtained by applying the step (0) of the procedure presented in Sect. 2.2 is reported in Table 2.

In practice, step (1) starts considering "c.age" (classes of age), and then adding the variables following the order presented in Table 2 until four of the available $X$ variables are cross-classified, as shown in Table 3. The selected combination is reported in bold. In fact, after adding "area5" (geographical macro regions) the

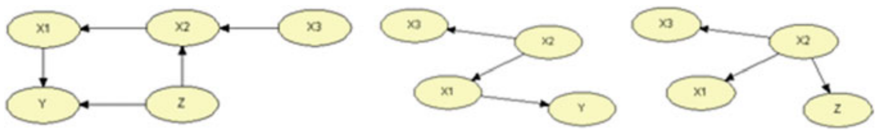


**Fig. 1** Simulated example: complete (*left*), $Y|X$ (*centre*) and $Z|X$ (*right*) models

**Table 1** Output of the procedure for selecting the matching variables

| $X$ variables | No. of $X$ | $\bar{n}$ | $\hat{d}$ |
|---|---|---|---|
| $X_1$ | 1 | 1250 | 0.1703 |
| $X_1 \times X_3$ | 2 | 625 | 0.1703 |
| $X_1 \times X_2 \times X_3$ | 3 | 312.5 | 0.1699 |

**Table 2** Initial ordering of $X$ variables. Step(0) of the algorithm

| | c.age | edu7 | marital | sex | hsize5 | area5 | urb |
|---|---|---|---|---|---|---|---|
| No. of categories | 5 | 7 | 3 | 2 | 5 | 5 | 3 |
| $\hat{d}$ | 0.0878 | 0.1056 | 0.1085 | 0.1097 | 0.1120 | 0.1133 | 0.1159 |
| Ranking | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

**Table 3** $X$ variables, average width $\hat{d}$ and average cell counts $\bar{n}$

| Combination of $X$ ($X_D$) | No. of $X$ | $\bar{n}$ | $\hat{d}$ |
|---|---|---|---|
| c.age | 1 | 86.0 | 0.0878 |
| c.age × sex | 2 | 43.0 | 0.0781 |
| c.age × sex × edu7 | 3 | 6.1 | 0.0714 |
| **c.age × sex × edu7 × area5** | **4** | **1.2** | **0.0608** |
| c.age × sex × edu7 × area5 × hsize5 | 5 | 0.2 | 0.0411 |

minimum of the average cell counts passes from 1.2 to 0.2, a value smaller than the decided stopping rule ($\bar{n} > 1$).

By comparing the results with those that would be obtained by exploring all the possible combinations of the $X$ variables (not reported here), it comes out that the procedure fails to identify the "best" model having four of the available $X$ variables, but the identified combination is the one that precedes immediately the best solution: "c.age × edu7 × area5 × hsize5" ($\hat{d} = 0.0575$).

The results confirm that the larger is the number of matching variables the lower is the uncertainty, but this reasoning is jeopardized by the fact that many matching variables increase the sparseness of the contingency tables; as can be seen in Table 3, passing from four to five $X$, the estimated average width of bounds shows a non-negligible decrease from 0.0608 to 0.0411, but the average cell frequency goes below 1 unit per cell.

## 4 Conclusions

The proposed procedure goes in the direction indicated by [4] avoiding separate analysis on the data sources at hand. The procedure is fully automatic and searches for the best combination of the available categorical common variables; it appears successful in identifying the various subsets of 1, 2, 3, etc. "best" matching variables. The stopping rule based on the sparseness of the tables meets the parsimony principle, however it is likely to be further refined to better catch the sparseness problem. On the other hand, further investigation is needed to understand whether the procedure can be further improved by introducing different methods for estimating (4) in presence of sparse tables (pseudo-Bayes estimator, etc.) and/or changing the way of measuring the overall uncertainty.

## References

1. Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, New York
2. Agresti A, Yang MC (1987) An empirical investigation of some effects of sparseness in contingency tables. Comput Stat Data Anal 5:9–21
3. Bishop YM, Fienberg SE, Holland PW (1975) Discrete Multivariate Analysis: Theory and Practice. MIT. Press, Cambridge, MA. Paperback edition
4. Cohen ML (1991) Statistical matching and microsimulation models. In: Citro, H (ed) Improving information for social policy decisions: The uses of microsimulation modeling, vol II Technical papers, Washington D.C
5. Conti PL, Marella D, Scanu M (2012) Uncertainty analysis in statistical matching. J Official Stat 28:69–88
6. D'Orazio M, Di Zio M, Scanu M (2006) Statistical matching: theory and practice. Wiley, Chichester
7. D'Orazio M (2016) StatMatch: statistical matching (aka data fusion). R package version 1.2.4 http://CRAN.R-project.org/package=StatMatch
8. Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York
9. Vantaggi B (2008) Statistical matching of multiple sources: a look through coherence. In J Approximate Reasoning 49:701–711

# Beyond Fuzzy, Possibilistic and Rough: An Investigation of Belief Functions in Clustering

**Thierry Denœux and Orakanya Kanjanatarakul**

**Abstract** In evidential clustering, uncertainty about the assignment of objects to clusters is represented by Dempster-Shafer mass functions. The resulting clustering structure, called a credal partition, is shown to be more general than hard, fuzzy, possibility and rough partitions, which are recovered as special cases. Different algorithms to generate a credal partition are reviewed. We also describe different ways in which a credal partition, such as produced by the EVCLUS or ECM algorithms, can be summarized into any of the simpler clustering structures.

## 1 Introduction

Clustering is one of the most important tasks in data analysis and machine learning. It aims at revealing some structure in a dataset, so as to highlight groups (clusters) of objects that are similar among themselves, and dissimilar to objects of other groups. Traditionally, we distinguish between *partitional* clustering, which aims at finding a partition of the objects, and *hierarchical* clustering, which finds a sequence of nested partitions.

Over the years, the notion of partitional clustering has been extended to several important variants, including fuzzy [1], possibilistic [2], rough [3, 4] and evidential clustering [5–7]. Contrary to classical (hard) partitional clustering, in which each object is assigned unambiguously and with full certainty to a cluster, these variants allow ambiguity, uncertainty or doubt in the assignment of objects to clusters. For this reason, they are referred to as *soft* clustering methods, in contrast with classical, *hard* clustering [8].

T. Denœux (✉)
Heudiasyc, Sorbonne Universités, Université de Technologie de Compiègne,
CNRS, UMR 7253, Paris, France
e-mail: Thierry.Denoeux@utc.fr

O. Kanjanatarakul
Faculty of Management Sciences, Chiang Mai Rajabhat University, Chiang Mai, Thailand
e-mail: orakanyaa@gmail.com

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_20

Among soft clustering paradigms, evidential clustering describes the uncertainty in the membership of objects to clusters using the formalism of *belief functions* [9]. The theory of belief functions is a very general formal framework for representing and reasoning with uncertainty. Roughly speaking, a belief function can be seen as a collection of sets with corresponding masses, or as a non additive measure generalizing a probability measure. Recently, evidential clustering has been successfully applied in various domains such as machine prognosis [10], medical image processing [11, 12] and analysis of social networks [13].

Because of its generality, the theory of belief functions occupies a central position among theories of uncertainty. The purpose of this paper is to show that, similarly, the evidential paradigm occupies a central position among soft clustering approaches. More specifically, we will show that hard, fuzzy, possibilistic and rough clustering can be all seen as special cases of evidential clustering. We will also study different ways in which a credal partition can be summarized into any of the other hard of soft clustering structures to provide the user with more synthetic views of the data.

The rest of this paper is structured as follows. In Sect. 2, the notion of credal partition will first be recalled, and algorithms to construct a credal partition will be reviewed. The relationships with other clustering paradigms will then be discussed in Sect. 3. Finally, Sect. 4 will conclude the paper.

## 2 Credal Partition

We first recall the notion of credal partition in Sect. 2.1. In Sect. 2.2, we briefly review the main algorithms for constructing credal partitions.

### 2.1 Credal Partition

Assume that we have a set $\mathcal{O} = \{o_1, \ldots, o_n\}$ of $n$ objects, each one belonging to one and only one of $c$ groups or clusters. Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ denote the set of clusters. If we know for sure which cluster each object belongs to, we can give a (hard) partition of the $n$ objects. Such a partition may be represented by binary variables $u_{ik}$ such that $u_{ik} = 1$ if object $o_i$ belongs to cluster $\omega_k$, and $u_{ik} = 0$ otherwise.

If objects cannot be assigned to clusters with certainty, then we can quantify cluster-membership uncertainty by mass functions $m_1, \ldots, m_n$, where each mass function $m_i$ is a mapping from $2^\Omega$ to $[0, 1]$, such that $\sum_{A\subseteq\Omega} m_i(A) = 1$. Each mass $m_i(A)$ is interpreted as a degree of support attached to the proposition "the true cluster of object $o_i$ is in $A$", and to no more specific proposition. A subset $A$ of $\Omega$ such that $m_i(A) > 0$ is called a *focal set* of $m_i$. The $n$-tuple $m_1, \ldots, m_n$ is called a *credal partition* [6].

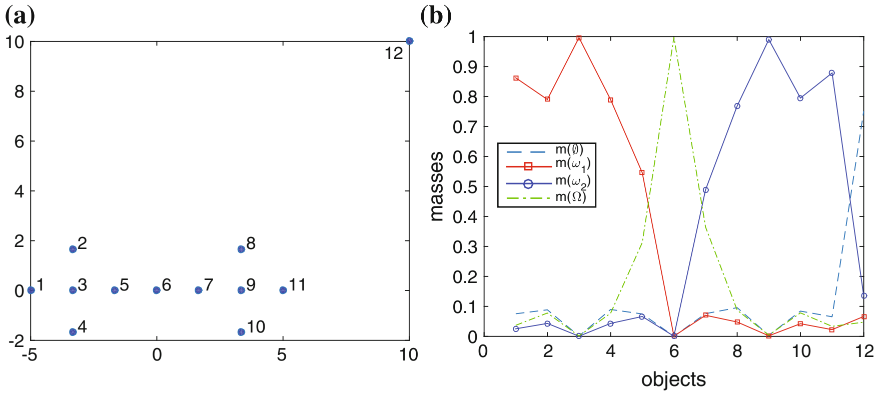


**Fig. 1** Butterfly dataset (**a**) and a credal partition (**b**)

*Example 1* Consider, for instance, the "Butterfly" dataset shown in Fig. 1(a). This dataset is adapted from the classical example by Windham [14], with an added outlier (point 12). Figure 1(b) shows the credal partition with $c = 2$ clusters produced by the Evidential $c$-means (ECM) algorithm [7]. In this figure, the masses $m_i(\emptyset)$, $m_i(\{\omega_1\})$, $m_i(\{\omega_2\})$ and $m_i(\Omega)$ are plotted as a function of $i$, for $i = 1, \ldots, 12$. We can see that $m_3(\{\omega_1\}) \approx 1$, which means that object $o_3$ almost certainly belongs to cluster $\omega_1$. Similarly, $m_9(\{\omega_2\}) \approx 1$, indicating almost certain assignment of object $o_9$ to cluster $\omega_2$. In contrast, objects $o_6$ and $o_{12}$ correspond to two different situations of maximum uncertainty: for object $o_6$, we have $m_6(\Omega) \approx 1$, which means that this object might as well belong to clusters $\omega_1$ and $\omega_2$. The situation is completely different for object $o_{12}$, for which the largest mass is assigned to the empty set, indicating that this object does not seem to belong to any of the two clusters.

## 2.2 *Evidential Clustering Algorithms*

Three main algorithms have been proposed to generate credal partitions:

1. The EVCLUS algorithm, introduced in [6], applies ideas from Multidimensional Scaling (MDS) [15] to clustering: given a dissimilarity matrix, it finds a credal partition such that the degrees of conflict between mass functions match the dissimilarities, dissimilar objects being represented by highly conflicting mass functions; this is achieved by iteratively minimizing a stress function. A variant of EVCLUS allowing one to use prior knowledge in the form of pairwise constraints was later introduced in [16], and several improvements to the original algorithm making it capable of handling large dissimilarity datasets have been reported in [17].

2. The Evidential $c$-means (ECM) algorithm [7] is a $c$-means-like algorithm that minimizes a cost function by searching alternatively the space of prototypes and the space of credal partitions. Unlike the hard and fuzzy $c$-means algorithms, ECM associates a prototype not only to clusters, but also to sets of clusters. The prototype associated to a set of clusters is defined as the barycenter of the prototypes of each single cluster in the set. The cost function to be minimized insures that objects close to a prototype have a high mass assigned to the corresponding set of clusters. A variant with adaptive metrics and pairwise constraints was introduced in [18], and a relational version for dissimilarity data (called RECM) has been proposed in [19].
3. The E$k$-NNclus algorithm [5] is a decision-directed clustering procedure based on the evidential $k$-nearest neighbor (E$K$-NN) rule [20]. Starting from an initial partition, the algorithm iteratively reassigns objects to clusters using the E$K$-NN rule, until a stable partition is obtained. After convergence, the cluster membership of each object is described by a Dempster-Shafer mass function assigning a mass to each cluster and to the whole set of clusters. The mass assigned to the set of clusters can be used to identify outliers. The procedure can be seen as searching for the most plausible partition of the data.

Each of these three algorithms have their strengths and limitations, and the choice of an algorithm depends on the problem at hand. Both ECM and E$K$-NN are very efficient for handling attribute data. E$K$-NN has the additional advantage that it can determine the number of clusters automatically, while EVCLUS and ECM produce more informative outputs (with masses assigned to any subsets of clusters). EVCLUS was shown to be very effective for dealing with non metric dissimilarity data, and the recent improvements reported in [17] make it suitable to handle very large datasets.

## 3 Relationships with Other Clustering Paradigms

In this section, we discuss the relationships between the notion of credal partition and other clustering structures. In Sect. 3.1, we show that hard, fuzzy, possibilistic and rough partitions are all special kinds of credal partitions. In Sect. 3.2, we describe how a general credal partition can be summarized in the form of any of the simpler structures mentioned previously.

### *3.1 Generality of the Notion of Credal Partition*

The notion of credal partition is very general, in the sense that it boils down to several alternative clustering structures when the mass functions composing the credal partition have some special forms (see Fig. 2).

Hard partition: If all mass functions $m_i$ are *certain* (i.e., have a single focal set, which is a singleton), then we have a hard partition, with $u_{ik} = 1$ if $m_i(\{\omega_k\}) = 1$, and $u_{ik} = 0$ otherwise.

Fuzzy partition: If the $m_i$ are *Bayesian* (i.e., they assign masses only to singletons, in which case the corresponding belief function becomes additive), then the credal partition is equivalent to a fuzzy partition; the degree of membership of object $i$ to cluster $k$ is $u_{ik} = m_i(\{\omega_k\})$.

Fuzzy partition with a noise cluster: A mass function $m$ such that each focal set is either a singleton, or the empty set may be called an *unnormalized Bayesian mass function*. If each mass function $m_i$ is unnormalized Bayesian, then we can define, as before, the membership degree of object $i$ to cluster $k$ a $u_{ik} = m_i(\{\omega_k\})$, but we now have $\sum_{k=1}^{c} u_{ik} \leq 1$, for $i = 1, \ldots, n$. We then have $m_i(\emptyset) = u_{i*} = 1 - \sum_{k=1}^{c} u_{ik}$, which can be interpreted as the degree of membership to a "noise cluster" [21].

Possibilistic partition: If the mass functions $m_i$ are *consonant* (i.e., if their focal sets are nested), then they are uniquely described by their contour functions

$$pl_i(\omega_k) = \sum_{A \subseteq \Omega, \omega_k \in A} m_i(A), \tag{1}$$

which are possibility distributions. We then have a possibilistic partition, with $u_{ik} = pl_i(\omega_k)$ for all $i$ and $k$. We note that $\max_k pl_i(\omega_k) = 1 - m_i(\emptyset)$.

Rough partition: Assume that each $m_i$ is *logical*, i.e., we have $m_i(A_i) = 1$ for some $A_i \subseteq \Omega$, $A_i \neq \emptyset$. We can then define the *lower approximation* of cluster $\omega_k$ as the set of objects that *surely* belong to $\omega_k$,

$$\omega_k^L = \{o_i \in \mathcal{O} | A_i = \{\omega_k\}\}, \tag{2}$$

and the *upper approximation* of cluster $\omega_k$ as the set of objects that *possibility* belong to $\omega_k$,

$$\omega_k^U = \{o_i \in \mathcal{O} | \omega_k \in A_i\}. \tag{3}$$

The membership values to the lower and upper approximations of cluster $\omega_k$ are then, respectively, $\underline{u}_{ik} = Bel_i(\{\omega_k\})$ and $\overline{u}_{ik} = Pl_i(\{\omega_k\})$. If we allow $A_i = \emptyset$ for some $i$, then we have $\overline{u}_{ik} = 0$ for all $k$, which means that object $o_i$ does not belong to the upper approximation of any cluster.

### 3.2 Summarization of a Credal Partition

A credal partition is a quite complex clustering structure, which often needs to be summarized in some way to become interpretable by the user. This can be achieved by transforming each of the mass functions in the credal partition into a simpler

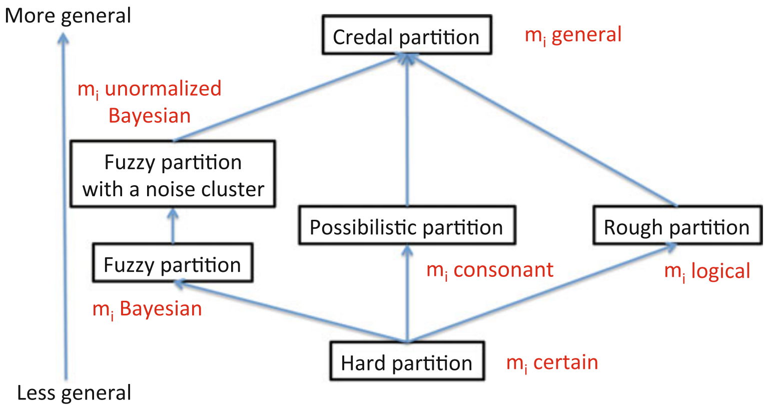


**Fig. 2** Relationship between credal partitions and other clustering structures

representation. Depending on the representation used, each of clustering structures mentioned in Sect. 3.1 can be recovered as different partial views of a credal partition. Some of the relevant transformations are discussed below.

Fuzzy and hard partitions: A fuzzy partition can be obtained by transforming each mass function $m_i$ into a probability distribution $p_i$ using the plausibility-probability transformation defined as

$$p_i(\omega_k) = \frac{pl_i(\omega_k)}{\sum_{\ell=1}^{c} pl_i(\omega_\ell)}, \quad k = 1, \ldots, c, \tag{4}$$

where $pl_i$ is the contour function associated to $m_i$, given by (1). By selecting, for each object, the cluster with maximum probability, we then get a hard partition.

Fuzzy partition with noise cluster: In the plausibility-probability transformation (4), the information contained in the masses $m_i(\emptyset)$ assigned to the empty set is lost. However, this information may be important if the dataset contains outliers. To keep track of it, we can define an unnormalized plausibility transformation as $\pi_i(\omega_k) = (1 - m_i(\emptyset))p_i(\omega_k)$, for $k = 1, \ldots, c$. The degree of membership of each object $i$ to cluster $k$ can then be defined as $u_{ik} = \pi_i(\omega_k)$ and the degree of membership to the noise cluster as $u_{i*} = m_i(\emptyset)$.

Possibilistic partition: A possibilistic partition can be obtained from a credal partition by computing a consonant approximation of each of the mass functions $m_i$ [22]. The simplest approach is to approximate $m_i$ by the consonant mass function with the same contour function, in which case the degree of possibility of object $o_i$ belonging to cluster $\omega_k$ is $u_{ik} = pl_i(\omega_k)$.

Rough partition: As explained in Sect. 3.1, a credal partition becomes equivalent to a rough partition when all mass functions $m_i$ are logical. A general credal partition can thus be transformed into a rough partition by deriving a set $A_i$ of clusters from each mass function $m_i$. This can be done either by selecting the focal set $A_{max} = \arg\max_{A \subseteq \Omega} m(A)$ with maximum mass as suggested in [7], or by the interval dominance decision rule

$$A^*(m_i) = \{\omega \in \Omega | \forall \omega' \in \Omega,\, pl_i^*(\omega) \geq m_i^*(\{\omega'\})\}, \tag{5}$$

where $pl_i^*$ and $m_i^*$ are defined, respectively, by $pl_i^* = pl_i/(1 - m_i(\emptyset))$ and $m_i^* = m_i/(1 - m_i(\emptyset))$. If the interval dominance rule is used, we may account for the mass assigned to the empty set by defining $A_i$ as follows,

$$A_i = \begin{cases} \emptyset & \text{if } m_i(\emptyset) = \max_{A \subseteq \Omega} m_i(A) \\ A^*(m_i) & \text{otherwise.} \end{cases} \tag{6}$$

## 4 Conclusions

The notion of credal partition, as well as its relationships with alternative clustering paradigms have been reviewed. Basically, each of the alternative partitional clustering structures (i.e., hard, fuzzy, possibilistic and rough partitions) correspond to a special form of the mass functions within a credal partition. We have also examined different ways in which a credal partition can be transformed into a simpler clustering structure for easier interpretation. As they build more complex clustering structures, credal clustering algorithms such as EVCLUS and ECM tend to be more computationally demanding than alternative algorithms. This issue can be dealt with by using efficient optimization algorithms and by restricting the form of the credal partition, making it possible to apply evidential clustering to large datasets with large numbers of clusters. First results along these lines have been reported in [17].

**Acknowledgments** This research was supported by the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" by the National Agency for Research (reference ANR-11-IDEX-0004-02).

## References

1. Bezdek JC (1981) Pattern Recognition with fuzzy objective function algorithm. Plenum Press, New-York
2. Krishnapuram R, Keller JM (1993) A possibilistic approach to clustering. IEEE Trans Fuzzy Syst 1:98–111
3. Lingras Pawan, Peters Georg (2012) Applying rough set concepts to clustering. In: Peters G, Lingras P, Ślezak D, Yao Y (eds) Rough Sets: Selected methods and applications in management and engineering. Springer, London, UK, pp 23–37
4. Peters Georg (2015) Is there any need for rough clustering? Pattern Recogn Lett 53:31–37
5. Denœux T, Kanjanatarakul O, Sriboonchitta S (2015) E$K$-NNclus: a clustering procedure based on the evidential $k$-nearest neighbor rule. Knowl-Based Syst 88:57–69

6. Denœux D, Masson MH (2004) EVCLUS: evidential clustering of proximity data. IEEE Trans Syst Man Cybern B, 34(1):95–109
7. Masson M-H, Denœux T (2008) ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recogn 41(4):1384–1397
8. Peters Georg, Crespo Fernando, Lingras Pawan, Weber Richard (2013) Soft clustering: fuzzy and rough approaches and their extensions and derivatives. Int J Approximate Reasoning 54(2):307–322
9. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton, N.J
10. Serir Lisa, Ramasso Emmanuel, Zerhouni Noureddine (2012) Evidential evolving Gustafson-Kessel algorithm for online data streams partitioning using belief function theory. Int J Approximate Reasoning 53(5):747–768
11. Benoît Lelandais Su, Ruan Thierry Denœux, Vera Pierre, Gardin Isabelle (2014) Fusion of multi-tracer PET images for dose painting. Med Image Anal 18(7):1247–1259
12. Makni Nasr, Betrouni Nacim, Colot Olivier (2014) Introducing spatial neighbourhood in evidential c-means for segmentation of multi-source images: Application to prostate multi-parametric MRI. Inf Fusion 19:61–72
13. Zhou Kuang, Martin Arnaud, Pan Quan, Liu Zhun-Ga (2015) Median evidential c-means algorithm and its application to community detection. Knowl-Based Syst 74:69–88
14. Windham MP (1985) Numerical classification of proximity data with assignment measures. J Classif 2:157–172
15. Borg I, Groenen P (1997) Modern multidimensional scaling. Springer, New-York
16. Antoine V, Quost B, Masson M-H, Denoeux T (2014) CEVCLUS: evidential clustering with instance-level constraints for relational data. Soft Comput 18(7):1321–1335
17. Denœux T, Sriboonchitta S, Kanjanatarakul, O (2016) Evidential clustering of large dissimilarity data. Knowledge-Based Syst 106:179–195
18. Antoine V, Quost B, Masson M-H, Denoeux T (2012) CECM: Constrained evidential c-means algorithm. Comput Stat Data Anal 56(4):894–914
19. Masson M-H, Denœux T (2009) RECM: relational evidential c-means algorithm. Pattern Recogn Lett 30:1015–1026
20. Denœux T (1995) A *k*-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans Syst Man Cybern 25(05):804–813
21. Davé RN (1991) Characterization and detection of noise in clustering. Pattern Recogn Lett 12:657–664
22. Dubois D, Prade H (1990) Consonant approximations of belief measures. Int J Approximate Reasoning 4:419–449

# Small Area Estimation in the Presence of Linkage Errors

**Loredana Di Consiglio and Tiziana Tuoto**

**Abstract** In Official Statistics, interest for data integration has been increasingly growing, though the effect of this procedure on the statistical analyses has been disregarded for a long time. In recent years, however, it is largely recognized that linkage is not an error-free procedure and linkage errors, as false links and missed links can invalidate standard estimates. More recently, growing attention is devoted to the effect of linkage errors on the subsequent analyses. For instance, Samart and Chambers (Samart in Aust N Z J Stat 56, 2014 [14]) consider the effect of linkage errors on mixed effect models. Their proposal finds a natural application in the context of longitudinal studies, where repeated measures are taken on the same individuals. In official statistics, the mixed models is largely exploited for small area estimation to increase detailed information at local level. In this work, an EBLUP estimator that takes account of the linkage errors is derived.

## 1 Data Integration and Impact of Linkage Errors

In Official Statistics, interest for data integration has been increasingly growing, though the effect of this procedure on the statistical analyses has been disregarded for a long time. In recent years, however, it is largely recognized that linkage is not an error-free procedure and linkage errors, as false links and missed links can invalidate standard estimates. More recently, growing attention is devoted to the effect of linkage errors on the subsequent analyses. Chambers [2] reviews the original work by Neter et al. [9] and its extensions by Scheuren and Winkler [15, 16] and by Lahiri and Larsen [8]. Moreover Chambers [2] suggests a Best Unbiased Estimator and its empirical version and proposes a maximum likelihood estimator with application to linear and logistic regression functions. An extension to sample-to-register linkage is also

L. Di Consiglio
Eurostat, Luxembourg, Germany
e-mail: loredana.di-consiglio@ec.europa.eu

T. Tuoto (✉)
Istat, Rome, Italy
e-mail: tuoto@istat.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_21

suggested. Samart and Chambers [14] consider the effect of linkage errors on mixed effect models, extending the settings in Chambers [2] and suggesting linkage errors adjusted estimators of variance effects under alternative methods. Their proposal finds a natural application in the context of longitudinal studies, where repeated measures are taken on the same individuals. In official statistics, the mixed models is largely exploited for small area estimation to increase detailed information at local level. Administrative data can be used to augment information collected on sample surveys, in order to expand auxiliary information and improve the model fitting for small area estimation. Linkage of external sources with basic statistical registers as well as with sample surveys can be carried out on different linkage scenarios. Di Consiglio and Tuoto [3] showed a sensitivity analysis for different alternative linkage error scenarios in the linear and logistic regression settings. In this work, we extend the analysis on the effects of linkage errors on the predictors based on a unit level mixed models for small area estimation when auxiliary variables are obtained through a linkage procedure with an external register. Under the assumption that false matches only occur within the same small area, the effect of linkage errors on small area predictors is given both by the effects on estimation of the fixed and random components, and by the effect on the variance matrix of the linked values and by the erroneous evaluation of covariates mean(s) on the set of sampled units (and consequently of unobserved population units). Following Chambers [2] in the sample-to-register linkage setting, in particular, assuming that sampling does not change the outcome of the linkage process, an EBLUP estimator based on the derived distribution of the linked values is obtained.

## 2 Linkage Model and Linkage Errors

The most widespread theory for record linkage is given by Fellegi and Sunter [5]. Given two lists (i.e. a register and a sample), say $L_1$ and $L_2$, of size $N_1$ and $N_2$, the linkage process can be viewed as a classification problem where the pairs in the cartesian product $\Omega = ((i,j), i \in L_1 \text{ and} j \in L_2)$ have to be assigned into two subsets $M$ and $U$, independent and mutually exclusive, such that M is the link set $(i = j)$ while U is the non-link set $(i \neq j)$. At the end of the linkage procedure, two kinds of errors may occur: the false match or false positive, when a pair is declared as a link but actually the two records are referred to different units, and the missing match or false negative, when the pair is declared as a non-link but actually the two records are referred to the same units. A good linkage strategy aims to minimize both probabilities of false match and missing match or, at least, to keep under assigned acceptable values. The probabilistic record linkage [5, 7] provides as output an evaluation of the probability of being a correct link given that the link is assigned:

$$\lambda_{ij} = m(\gamma_{ij})P(M^*)/(m(\gamma_{ij})P(M^*) + u(\gamma_{ij})P(U^*)), \tag{1}$$

where $m(\gamma_{ij})$ is the conditional probability of the comparison vector $\gamma$ between $i \in L_1$ and $j \in L_2$ given that the pair belongs to set $M$ and $u(\gamma_{ij})$ is the conditional probability of the comparison vector $\gamma$ given that the pair belongs to set $U$. These quantities $\lambda_{ij}$ will be exploited for adjusting the linkage errors in the small area estimation framework described in the next section.

# 3 Small Area Estimation Based on Unit Linear Mixed Model

When sample size is not big enough, the standard estimators are often not reliable enough to produce estimators at finer level of (geographical) detail (for a review, see [10]). The EBLUP estimators based on a unit level models was firstly proposed by Battese et al. [1] to improve the reliability of estimators exploiting the relationship between the target variable and external auxiliary variables.

## *3.1 The Unit Linear Mixed Model*

Let us suppose that the population units can be grouped in $D$ domains, let $Y$ be the target variable and $X$ auxiliary variables observed on the same units. Let us assume a linear mixed relationship between the target variable and the covariates

$$y_{id} = X_{id}^T\beta + u_d + e_{id}, \ \ i = 1, \dots, N_d, \ \ d = 1, \dots, D, \tag{2}$$

where $\beta$ is a $p$-dimensional vector of fixed regression coefficients and $u_d$, $d = 1, \dots, D$, are the i.i.d. random variables related to the specific or domain contributions, with $E(u_d) = 0$ and $V(u_d) = \sigma_u^2$ and independent errors $e_{id}$ i.i.d. with $E(e_{id}) = 0$ and $V(e_{id}) = \sigma_e^2$. In matrix notation

$$Y = X\beta + Zu + e$$

where $Z$ is the area design matrix, $Z = Blockdiag(Z_d = 1_{N_d}; d = 1 \cdots D)$. The total variance is then $V(Y) = V = \sigma_u^2 ZZ^T + \sigma_e^2 I$ or $V = diag(V_d; d = 1 \cdots D)$ with $V_d = \sigma_e^2 I_{N_d} + \sigma_u^2 Z_d Z_d^T$. When $\sigma_u^2$ and $\sigma_e^2$ are known, the BLUP estimator of a small area mean or totals $\bar{Y}_d$, is given by

$$\hat{\bar{Y}}_d^{BLUP} = \frac{1}{N_d}\left(\sum_{i \in s_d} y_{id} + \sum_{i \in s_d^c} \hat{y}_{id}^{BLUP}\right) \tag{3}$$

where $\hat{y}_{id}^{BLUP} = X_{id}^T\hat{\beta} + \hat{u}_d$ with

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

and $\hat{u} = \sigma_u Z^T V(y - X\hat{\beta})$. An EBLUP is obtained by plugging the estimates $\widehat{sigma}_u$ and $\hat{\sigma}_e$ in the previous expressions, (see Sect. 3.3).

### 3.2 The Unit Linear Mixed Model Under RL

When the covariates $X$ and target variable $Y$ are not observed on the same set of data, but are obtained for example by linking a sample with a register, the use of the previous relationships on the observed data may produce biased estimates. Following [2] and [14], for unit $i$ let $y_{id}^*$ be the value of the variable matched with the value $X_{id}$. Let $Z_2$ be a blocking variable, measured without error on both the Y-register and the $X$-register, that partitions both registers so that all linkage errors occur within the groups of records defined by the distinct values of this variable. An exchangeable linkage errors model can be defined by assuming that the probability of correct linkage is the same for all records in a block, q, $q = 1, \dots, Q$. Under the following standard assumptions [2]:

1. the linkage is complete, i.e. the $X$-register and Y-register refer to the same population and have no duplicates, so the smallest Y-register is contained in the biggest $X$-register
2. the linkage is one to one between the $Y$- and $X$-registers
3. exchangeable linkage errors model, see [2]

then, omitting the blocking index q for simplicity of notation, the observed variable is a permutation of the true one $Y^* = AY$ where $A$ is a random permutation matrix such that $E(A|X) = E$. Being $Pr(a_{ii} = 1|X) = Pr(correct\ linkage) = \lambda$ and $Pr(a_{ij} = 1|X) = Pr(uncorrect\ linkage) = \gamma$ then the expected value can be written as:

$$E = (\lambda\gamma)I + \gamma 1 1^T \tag{4}$$

Samart and Chambers [14] proposed a ratio type corrected estimator for $\beta$

$$\tilde{\beta}_R = (X^T V^{-1} E X)^{-1} X^T V^{-1} y^* \tag{5}$$

Furthermore, by exploiting the relationship between the variable $y^*$ and X a BLUE can be obtained as

$$\tilde{\beta}_{BLUE} = (X^T E^T \Sigma E X)^{-1} X^T E^T \Sigma y^* \tag{6}$$

by taking into account the derived variance of the observed $y^*$

$$V(Y^*) = \Sigma^{-1} = \sigma_u^2 K + \sigma_e^2 I + V \tag{7}$$

where

$$V \approx diag((1-\lambda)(\lambda(f_i - \bar{f}) + \bar{f}^{(2)} - \bar{f}^2)) \tag{8}$$

being $f_i = X_i\hat{\beta}$ and $K$ a function of the number of areas within a block, block-group sizes and $\lambda s$ (see [14]).

### 3.3 Estimation of Variance Components (ML)

As $\sigma_u$ and $\sigma_e$ are unknown, they have to be estimated; common methods are the methods of moments, ML or REML [6, 17]. Here we restrict to the ML, assuming multivariate normal distribution. In general, there are no analytical expressions for the variance component estimators obtained by using ML. Samart and Chambers [14] use the method of scoring as an algorithm to obtain the estimators. In the standard case where the variables are recorded on the sample, we have $y \sim N(X\beta; V)$ For the record linkage case, recall that $y^* \sim N(Ef; \Sigma)$. The scoring algorithm can be applied on the derivatives of the previous likelihood. Estimation of $\beta$ is then obtained by replacing the variance components estimates and clearly an iterative process is needed.

### 3.4 Small Area Estimation Under Linkage Errors

For the purpose of small area estimation, the scenario to be considered is the linkage of a sample with a register. Here we assume that the register is complete, i.e. neither duplicates and coverage issues occur. This setting is considered in Chambers [2] when the second data set is included in the first one. Following the proposed framework, we also assume that the record linkage process is independent of the sampling process. Chambers [2] assumes that an hypothetical linkage can be performed before the sampling process. Under these conditions, the matrices $E$, $V$ and $\Sigma$ depend only on the blocking variables and linkage errors, so the use of sampling weights is not needed. Besides these assumptions, as specified in Sect. 3.2 we assume an exchangeable linkage errors model, i.e. the linkage errors occur only within the same block where records have the same probability of being correctly linked. Finally, we assume that small area coincides with blocks. In this case, of course,

$$\hat{\bar{Y}}^* = \hat{\bar{Y}}$$

and we can exploit the distribution of $Y^*$ to obtain the EBLUP estimator:

$$\hat{\bar{Y}}_d^{*BLUP} = \frac{1}{N_d}\left(\sum_{i\in s_d} y_{id}^* + \sum_{i\in s_d^c} \hat{y}_{id}^{BLUP}\right) \tag{9}$$

where $\hat{y}_{id}^{BLUP} = EX\tilde{\beta}_{BLUE} + \hat{u}_d$ and $\hat{u} = \sigma_u Z^T \Sigma^{-1}(y^* - EX\tilde{\beta}_{BLUE})$. For computational ease the sum of the predicted values of non sampled units can be obtained as the difference of the total population predicted values and sum of the sample predicted values. Note that given the assumptions, the population matrix E is known. The EBLUP estimators are given by replacing in (9) the obtained estimators of regression coefficients and variance components.

## 4 Results on Simulated Data

From the fictitious population census data [4] created for the ESSnet DI, which was an European project on data integration run from 2009 to 2011, two different populations A and B were created on the basis of the following linear mixed models

A: $X \sim [1, Uniform(0,1)]$; $\beta = [2,4]$; $u \sim N(0,1)$; $e \sim N(0,3)$; Realized $Var(u) = 0.60$

B: $X \sim [1, Uniform(0,1)]$; $\beta = [2,4]$; $u \sim N(0,3)$; $e \sim N(0,1)$; Realized $Var(u) = 0.65$.

The population size is over 20000 records for both A and B; they also contain linking variables (names, dates of birth, addresses) for individual identification with missing values and typos, mimicking real situation. The small domains are defined as aggregation of postal codes, assigning 18 areas. For each population, 100 replicated samples of size 1000 were independently randomly selected without replacement; finally on each replicated setting, the sample containing the Y variable was linked with the register reporting the X variables, represented by the two populations. The linkage was performed by means of the batch version of the software RELAIS [11] implementing the probabilistic record linkage model [5, 7]. Two different scenarios were considered, characterized by two different sets of linking variables: in Scenario 1 we use Day, Month and Year of Birth; in Scenario 2 we adopt Day and Year of Birth and Gender. In the first scenario, variables with the higher identifying power were used for the linkage with respect to the second one, producing less linkage errors (in terms of both missing and false links) affecting the results. Table 1 summaries the results of the linkage procedures for the 100 replications, illustrating the number of declared matches (average) and statistics for the probability of false link $\lambda$ and the probability of missing link.

**Table 1** Linking results

| Scenario | Declared matches[a] | Min ($\lambda$) | Mean ($\lambda$) | Max ($\lambda$) | Min (prob of missing link) |
|---|---|---|---|---|---|
| 1 | 937 | 0.000 | 0.028 | 0.25 | 0.063 |
| 2 | 957 | 0.000 | 0.14 | 0.357 | 0.043 |

[a]average values in 100 replications

**Table 2** Average relative absolute differences of benchmark estimates and those obtained by estimators C and D

| Scenario | POP A | | POP B | |
|---|---|---|---|---|
| | Estimates C | Estimates D | Estimates C | Estimates D |
| 1 | 0.005 | 0.004 | 0.002 | 0.001 |
| 2 | 0.007 | 0.007 | 0.006 | 0.002 |

In the simulation, four estimators are considered for comparison:

(A) the EBLUP with X and Y observed on the same dataset, i.e. no linkage is assumed in this setting
(B) the EBLUP on the subset of linked records, in this setting we reduce the sample size to the linked record but we do not introduce linkage errors; this is our benchmark.
(C) the naïve EBLUP on the subset of linked records, considering X and Y observed on two different dataset (without adjustment error linkage)
(D) the adjusted EBLUP estimator

In Table 2 the average relative absolute difference is reported. As it is apparent in our scenarios the EBLUP is not on average very sensitive to the resulting linkage errors, however the adjusted estimator always improves the naïve estimator. The regression coefficients and the variance components estimates are not reported but the improvement is in the same direction.

## 5 Concluding Remarks and Future Works

We have examined the possibility to adjust the EBLUP for small area estimation based on a unit level mixed model when the auxiliary variables come from a register that has to be linked with the sample reporting the target variable. The proposal produces a slight improvement, when the magnitude of linkage errors is relatively low (in the worst scenario, the average in areas and replications is less than 15 %). One can expect a more sensitive improvement with higher linkage error levels. The proposed adjustment is still subject to very restrictive assumptions, such as the exchangeability of linkage errors, the small areas coincident to blocks for the linkage process and finally the assumption of known linkage errors. In presence of estimation of the latter ones, the bias-variance trade-off of the adjustment should be assessed.

## References

1. Battese GE, Harter RM, Fuller WA (1988) An error-components model for prediction of crop areas using survey and satellite data. J Am Stat Assoc 83:28–36
2. Chambers R (2009) Regression analysis of probability-linked data. In: Official Statistics Research Series, vol 4
3. Di Consiglio L, Tuoto T (2014) When adjusting for bias due to linkage errors: a sensitivity analysis. In: European Conference on Quality in Official Statistics (Q2014), Vienna, 3–5 June 2014
4. Essnet DI- McLeod, Heasman, Forbes (2011) Simulated data for the on the job training. http://www.cros-portal.eu/content/job-training
5. Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64:1183–1210
6. Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–338
7. Jaro M (1989) Advances in record linkage methodology as applied to matching the 1985 test census of Tampa, Florida. J Am Stat Assoc 84:414–420
8. Lahiri P, Larsen MD (2005) Regression analysis with linked data. J Am Stat Assoc 100:222–230
9. Neter J, Maynes ES, Ramanathan R (1965) The effect of mismatching on the measurement of response errors. J Am Stat Assoc 60:1005–1027
10. Rao JNK (2003) Small area estimation. Wiley, New York
11. RELAIS 3.0 Users Guide (2015). http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/relais
12. Samart K (2011) Analysis of probabilistically linked data. Ph.D. thesis, School of Mathematics and Applied Statistics, University of Wollongong
13. Samart K, Chambers R (2010) Fitting linear mixed models using linked data, centre for statistical and survey methodology. University of Wollongong, Working Paper 18-10
14. Samart K, Chambers R (2014) Linear regression with nested errors using probability-linked data. Aust N Z J Stat 56
15. Scheuren F, Winkler WE (1993) Regression analysis of data files that are computer matched Part I. Surv Methodol 19:39–58
16. Scheuren F, Winkler WE (1997) Regression analysis of data files that are computer matched-part II. Surv Methodol 23:157–165
17. Searle SR, Casella G, McCulloch CE (2006) Variance components. Wiley, New York
18. Tancredi A, Liseo B (2011) A hierachical Bayesian approach to record linkage and population size problems. Ann Appl Stat 5:1553–1585

# A Test for Truncation Invariant Dependence

F. Marta L. Di Lascio, Fabrizio Durante and Piotr Jaworski

**Abstract** A test is proposed to check whether a random sample comes from a truncation invariant copula $C$, that is, if $C$ is the copula of a pair $(U, V)$ of random variables uniformly distributed on $[0, 1]$, then $C$ is also the copula of the conditional distribution function of $(U, V \mid U \leq \alpha)$ for every $\alpha \in (0, 1]$. The asymptotic normality of the test statistics is shown. Moreover, a procedure is described to simplify the approximation of the asymptotic variance of the test. Its performance is investigated in a simulation study.

## 1 Introduction

Let $(X, Y)$ be a random pair describing a phenomenon of interest. To obtain a parametric model for the joint distribution function $H$ of $(X, Y)$, a frequently used starting point is Sklar's recipe, which states that $H$ can be expressed as

$$H(x, y) = C(F(x), G(y)) \qquad \text{for all } (x, y) \in \mathbb{R}^2, \tag{1}$$

in terms of a unique bivariate copula $C$ and the univariate margins $F$ and $G$. A copula model is hence obtained when suitable univariate distribution functions $F$ and $G$ are chosen in (1), and a copula is selected from a specific family $\mathcal{F}$.

Therefore, it is of interest in many applications to check whether the unknown copula $C$ belongs to the given class $\mathcal{F}$. In the literature, several tests of this type have been developed for $\mathcal{F}$ being the family of Archimedean copulas [3, 13], extreme-value copulas [2, 9, 16], or for other classes of copulas with special dependence

F.M.L. Di Lascio · F. Durante
Free University of Bozen-Bolzano, Bolzano, Italy
e-mail: marta.dilascio@unibz.it

F. Durante
e-mail: fabrizio.durante@unibz.it

P. Jaworski (✉)
University of Warsaw, Warsaw, Poland
e-mail: p.jaworski@mimuw.edu.pl

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_22

properties [1, 11, 12]. Here, we consider a novel family of copulas introduced in [5] (see also [4, 6, 14, 15]). Copulas belonging to this class, here denoted by $\mathcal{C}^{\text{LT}}$, are characterized in terms of left truncation invariant property, i.e. if a random pair $(U, V)$ is distributed according to a copula $C \in \mathcal{C}^{\text{LT}}$, then $C$ is also the copula related to the conditional distribution function of $(U, V \mid U \leq \alpha)$ for every $\alpha \in (0, 1]$.

Specifically, our main purpose is to derive a procedure to test the null hypothesis $H_0\colon C \in \mathcal{C}^{\text{LT}}$ against the alternative $H_1\colon C \notin \mathcal{C}^{\text{LT}}$. As for many goodness-of-fit tests reviewed by [7, 8], the proposed procedure is based on pseudo-observations. As known, such approach is justified because, the pseudo-observations, similarly as copulas, are invariant under strictly increasing transformations of $X$ and $Y$.

The manuscript is organized as follows. Section 2 presents the main testing procedure (as described in [4]) and discusses the asymptotic normality of the test. Section 3 presents a way to approximate the variance of the test statistics.

## 2 The Testing Procedure

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a random pair $(X, Y)$ with distribution function $H$ with unknown continuous margins $F$ and $G$, and unknown copula $C$. Let $F_n$ and $H_n$ be the empirical distribution functions given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x), \qquad H_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x, Y_i \leq y) \tag{2}$$

for all $x, y \in \mathbb{R}$.

We set $I = \mathbb{E}\,(H(X, Y))$ and $J = \mathbb{E}\,(F(X)H(X, Y))$. Moreover, we consider their empirical counterparts

$$I_n = \frac{1}{n}\sum_{i=1}^{n} H_n(X_i, Y_i), \quad J_n = \frac{1}{n}\sum_{i=1}^{n} F_n(X_i)H_n(X_i, Y_i). \tag{3}$$

Here we present a test in order to check

$$H_0\colon C \in \mathcal{C}^{\text{LT}} \quad \text{versus} \quad H_1\colon C \notin \mathcal{C}^{\text{LT}},$$

Now, as a consequence of the results in [4], if $(X, Y)$ is distributed according to $H = C(F, G)$, then, under the null hypothesis that $C \in \mathcal{C}^{\text{LT}}$, the vector

$$R_H(X, Y) = \left(F(X), \frac{C(F(X), G(Y))}{F(X)}\right)$$

is formed by independent components. Thus, in order to perform a test for the null hypothesis of interest, one can consider the following null hypothesis

$$H_0^*: \ R_H(X,Y) \text{ is formed by independent components.}$$

As $H_0$ implies $H_0^*$, we reject $H_0$ if $H_0^*$ is rejected. Following [4], the test for the null hypothesis $H_0^*$ can be performed by considering the test statistic

$$T_n = \sqrt{n}\left(\frac{3}{2}J_n - I_n\right). \tag{4}$$

In [4], it is shown that, under $H_0$ and suitable regularity assumptions, for every $\delta > 0$

$$\lim_{n\to\infty} \mathbb{P}(|T_n| < \delta) = 0,$$

i.e. the test statistics is consistent.

Moreover, it can be also shown that the test statistics is also asymptotically normal. Indeed, suppose that the arrow $\rightsquigarrow$ denotes weak convergence in the sense of [17]. Under suitable regularity conditions, it can be proved that $\mathbb{F}_n = \sqrt{n}(F_n - F) \rightsquigarrow \mathbb{F}$, where $\mathbb{F}(x) = \beta \circ F(x)$ and $\beta$ is a Brownian bridge. Also, $\mathbb{H}_n = \sqrt{n}(H_n - H) \rightsquigarrow \mathbb{H}$, where $\mathbb{H}(x,y) = \mathbb{C}(F(x), G(y))$ and $\mathbb{C}$ is a $C$-Brownian bridge. Moreover, both processes are centered Gaussian processes.

From [10, Sect. 3.5], it follows that $\sqrt{n}(I_n - I)$ and $\sqrt{n}(J_n - J)$ jointly converge to

$$\mathbb{I} = \mathbb{Z} + \int \mathbb{H}(x,y)dH(x,y), \tag{5}$$

and

$$\mathbb{J} = \mathbb{W} + \int \mathbb{F}(x)H(x,y)dH(x,y) + \int F(x)\mathbb{H}(x,y)dH(x,y), \tag{6}$$

where $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{H(X_i,Y_i) - I\} \rightsquigarrow \mathbb{Z}$ and $W_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{F(X_i)H(X_i, Y_i) - J\} \rightsquigarrow \mathbb{W}$. Finally, from [10, Sect. 3.5] the test statistics $T_n$ is approximately Gaussian for large $n$.

Thus, in order to determine the rejection and non-rejection regions of the test, it would be enough to approximate the variance of $T_n$, which depends however on the (unknown) copula. To overcome this problem, here we propose to calculate the variance of $T_n$ from a reference type of parametric family of copulas and, then, apply it to the general class. Admittedly, this is just selecting one specific family out of the whole class $\mathcal{C}^{\mathrm{LT}}$ under the null, but the selection shows reasonable performances, as we will show. A natural candidate for such a reference distribution is the Clayton family of copulas, as discussed in [4].

Under the previous setting, the testing procedure (at the significance level $\alpha$) goes as follows.

1. Given the i.i.d. observations $(x_1, y_1), \dots, (x_n, y_n)$ from $(X, Y)$, calculate the corresponding empirical Kendall's $\widehat{\tau}_n$ and the value of the test statistics $\widehat{T}_n$ from Eq. (4).

2. Consider the approximate standard deviation $\widehat{\sigma}_n$ of the test statistics obtained from the Clayton copula with Kendall's $\tau$ equal to $\widehat{\tau}_n$. (In the next section we will discuss how to approximate $\widehat{\sigma}_n$.)
3. Reject $H_0 \colon C \in \mathcal{C}^{\mathrm{LT}}$ if

$$\left| \frac{\widehat{T}_n}{\widehat{\sigma}_n} \right| > \phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where $\phi$ denotes the standard Gaussian cumulative distribution function.

The performance of the whole procedure is illustrated in a small simulation study in Tables 1 and 2. As it can be seen, the performance is generally acceptable, even though the procedure is more restrictive than the results in [4]. However, notice that this latter procedure is more expensive in terms of computational complexity.

**Table 1** Rejection percentage (over $N = 1000$ replications) of $H_0 \colon C \in \mathcal{C}^{\mathrm{LT}}$ (at different significance levels $\alpha$) for a random sample of size $n = 200$ generated from different copulas with a specified Kendall's $\tau$

| True copula | $\alpha = 0.01$ | | | $\alpha = 0.05$ | | | $\alpha = 0.1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\tau =$ 0.25 | $\tau =$ 0.5 | $\tau =$ 0.75 | $\tau =$ 0.25 | $\tau =$ 0.5 | $\tau =$ 0.75 | $\tau =$ 0.25 | $\tau =$ 0.5 | $\tau =$ 0.75 |
| Clayton | 0.028 | 0.029 | 0.056 | 0.121 | 0.140 | 0.195 | 0.234 | 0.249 | 0.316 |
| Frank$^{\mathrm{LT}}$ | 0.032 | 0.054 | 0.070 | 0.125 | 0.158 | 0.198 | 0.242 | 0.265 | 0.306 |
| Surv Clayton | 0.921 | 1.000 | 1.000 | 0.997 | 1.000 | 1.000 | 0.990 | 1.000 | 1.000 |
| Surv Gumbel | 0.041 | 0.231 | 0.695 | 0.135 | 0.483 | 0.886 | 0.238 | 0.628 | 0.952 |
| Gumbel | 0.727 | 0.999 | 1.000 | 0.884 | 1.000 | 1.000 | 0.935 | 1.000 | 1.000 |
| Frank | 0.304 | 0.920 | 0.997 | 0.551 | 0.982 | 0.999 | 0.680 | 0.994 | 1.000 |
| Gaussian | 0.288 | 0.923 | 0.999 | 0.540 | 0.975 | 1.000 | 0.702 | 0.991 | 1.000 |
| $t$-Student | 0.250 | 0.846 | 0.993 | 0.470 | 0.953 | 1.000 | 0.621 | 0.976 | 1.000 |

**Table 2** Rejection percentage (over $N = 1000$ replications) of $H_0 \colon C \in \mathcal{C}^{\mathrm{LT}}$ (at different significance levels $\alpha$) for a random sample of size $n = 200$ generated from different copulas with a specified Kendall's $\tau$

| True copula | $\alpha = 0.01$ | | | $\alpha = 0.05$ | | | $\alpha = 0.1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\tau =$ $-0.25$ | $\tau =$ $-0.5$ | $\tau =$ $-0.75$ | $\tau =$ $-0.25$ | $\tau =$ $-0.5$ | $\tau =$ $-0.75$ | $\tau =$ $-0.25$ | $\tau =$ $-0.5$ | $\tau =$ $-0.75$ |
| Clayton | 0.030 | 0.030 | 0.029 | 0.118 | 0.120 | 0.122 | 0.227 | 0.231 | 0.251 |
| Cl$^{0,1}$ | 0.035 | 0.035 | 0.037 | 0.127 | 0.126 | 0.139 | 0.231 | 0.245 | 0.234 |
| Frank | 0.487 | 0.964 | 0.997 | 0.738 | 0.995 | 1.000 | 0.842 | 0.999 | 1.000 |
| Gaussian | 0.453 | 0.959 | 0.999 | 0.747 | 0.997 | 0.998 | 0.999 | 1.000 | 1.000 |
| $t$-Student | 0.392 | 0.905 | 0.990 | 0.657 | 0.978 | 1.000 | 0.781 | 0.994 | 1.000 |

## 3 Approximation of the Variance of the Test Statistics

In order to approximate the variance of the estimator $T_n$ under the null hypothesis that $C \in \mathcal{C}^{\mathrm{LT}}$, we proceed as follows.

First, if we replace $\mathbb{H}$ by $\mathbb{H}_n$ and $\mathbb{F}$ by $\mathbb{F}_n$ in Eqs. (5) and (6), one ends up with

$$\int \mathbb{H}(x,y)dH(x,y) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\bar{H}(X_i,Y_i) - I\} + o_P(1),$$

with $\bar{H}(x,y) = \mathbb{P}(X \geq x, Y \geq y)$ being the survival function associated with $H$. Similarly,

$$\int \mathbb{F}(x)H(x,y)dH(x,y) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi(X_i) - J\} + o_P(1),$$

with $\phi(x) = \mathbb{E}\left(H(X,Y)\mathbb{1}(X \geq x)\right)$, and

$$\int \mathbb{H}(x,y)F(x)dH(x,y) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\psi(X_i) - J\} + o_P(1),$$

with $\psi(x,y) = \mathbb{E}\left(F(X)\mathbb{1}(X \geq x, Y \geq y)\right)$.

As a result, $\sqrt{n}(I_n - I)$ and $\sqrt{n}(J_n - J)$ have the same asymptotic distribution as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{H(X_i,Y_i) + \bar{H}(X_i,Y_i) - 2I\}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{F(X_i)H(X_i,Y_i) + \phi(X_i) + \psi(X_i,Y_i) - 3J\},$$

respectively.

In particular, if we consider the pseudo-observations $(U,V) = (F(X),G(Y))$, the asymptotic variance of $T_n$ can be approximated from the variance of

$$\begin{aligned}
&\frac{3}{2}\left(F(X)H(X,Y) + \phi(X) + \psi(X,Y) - 3J\right) - \big(H(X,Y) + \bar{H}(X,Y) - 2I\big)\\
=&\frac{3}{2}\big(UC(U,V) + \phi(F^{-1}(U)) + \psi(F^{-1}(U), G^{-1}(V)) - 3J\big)\\
&-\big(C(U,V) + \bar{C}(U,V) - 2I\big),
\end{aligned}$$

where $\bar{C}$ is the survival function associated with $C$.

Assume now that $C \in \mathcal{C}^{\mathrm{LT}}$. As a consequence of [4, Corollary 3.1], if $(U, V) \sim C$, then $U$ and $Z = \frac{C(U,V)}{U}$ are independent. In such a case, it follows that

$$\begin{aligned}\phi(F^{-1}(u)) &= \mathbb{E}(C(U, V))\mathbb{1}(U \geq u)) = \mathbb{E}(UZ\mathbb{1}(U \geq u)) \\ &= \frac{1}{2}(1-u^2)\mathbb{E}(Z) = (1-u^2)\mathbb{E}(C(U, V)) = (1-u^2)I.\end{aligned}$$

Analogously,

$$\begin{aligned}\psi(F^{-1}(u), G^{-1}(v)) &= \mathbb{E}\left(F(X)\mathbb{1}\left(X \geq F^{-1}(u), Y \geq G^{-1}(v)\right)\right) \\ &= \mathbb{E}\left(U\mathbb{1}\left(U \geq u, V \geq v\right)\right) \\ &= \mathbb{E}\left(U\mathbb{1}\left(U \geq u, Z \geq \frac{C(U, v)}{U}\right)\right) \\ &= \mathbb{E}\left(U\mathbb{1}\left(U \geq u\right)\mathbb{E}\left(\mathbb{1}\left(Z \geq \frac{C(U, v)}{U}\right)\Big|U\right)\right) \\ &= \frac{1}{2}(1-u^2) - \mathbb{E}\left(U\mathbb{1}\left(U \geq u\right)F_Z\left(\frac{C(U, v)}{U}\right)\right) \\ &= \frac{1}{2}(1-u^2) - \mathbb{E}\left(U\mathbb{1}\left(U \geq u\right)\partial_1 C(U, v)\right),\end{aligned}$$

where the last equality follows from the properties of the distribution function $F_Z(z) = \mathbb{P}(Z \leq z)$ given in [4, Theorem 3.1]. Thus

$$\begin{aligned}\psi(F^{-1}(u), G^{-1}(v)) &= \frac{1}{2}(1-u^2) - \int_u^1 \xi\partial_1 C(\xi, v)d\xi \\ &= \frac{1}{2}(1-u^2) - v + uC(u, v) + \int_u^1 C(\xi, v)d\xi,\end{aligned}$$

where the last expression is obtained using integration by parts. Thus, taking into account the previous equalities and $3J = 2I$, the asymptotic variance of $T_n$ can be estimated from the variance of

$$3UC(U, V) - 2C(U, V) + \left(-\frac{3}{4}U^2 + U - \frac{3IU^2}{2}\right) - \frac{V}{2} + \frac{I}{2} - \frac{1}{4} + \frac{3}{2}\int_U^1 C(\xi, V)d\xi. \tag{7}$$

For example, if $C$ is the independence copula, then the standard deviation of the previous expression is easily verified to be equal to $\sqrt{10}/60$. For bivariate Clayton copulas $C_\theta$ with Kendall's $\tau$ spanning from $-0.99$ to $0.99$ the standard deviation is shown in Table 3 (values obtained by Monte-Carlo procedures with $10^7$ replications).

**Table 3** Standard deviation (SD) derived from Eq. (7) for a bivariate Clayton copula $C_\theta$ with different parameter values

| $\theta$ | $\tau$ | SD |
|---|---|---|
| −0.9950 | −0.99 | 0.0091 |
| −0.9744 | −0.95 | 0.0198 |
| −0.9189 | −0.85 | 0.0323 |
| −0.8571 | −0.75 | 0.0394 |
| −0.7879 | −0.65 | 0.0440 |
| −0.6667 | −0.50 | 0.0486 |
| −0.5185 | −0.35 | 0.0517 |
| −0.4000 | −0.25 | 0.0530 |
| −0.2609 | −0.15 | 0.0535 |
| −0.0952 | −0.05 | 0.0532 |
| 0.1053 | 0.05 | 0.0521 |
| 0.3529 | 0.15 | 0.0503 |
| 0.6667 | 0.25 | 0.0480 |
| 1.0769 | 0.35 | 0.0450 |
| 2.0000 | 0.50 | 0.0390 |
| 3.7143 | 0.65 | 0.0308 |
| 6.0000 | 0.75 | 0.0239 |
| 11.3333 | 0.85 | 0.0157 |
| 38.0000 | 0.95 | 0.0057 |
| 198.0000 | 0.99 | 0.0026 |

**Acknowledgments** We would like to thank Bruno Rémillard for many useful comments about the asymptotic behavior of our test procedure. The authors have been supported by Free University of Bozen-Bolzano, Italy, via the projects COCCO and MODEX. The third author acknowledges the support from National Science Centre, Poland, under project 2015/17/B/HS4/00911.

## References

1. Berghaus B, Bücher A (2014) Nonparametric tests for tail monotonicity. J Econom 180(2):117–126
2. Bücher A, Dette H, Volgushev S (2011) New estimators of the Pickands dependence function and a test for extreme-value dependence. Ann Stat 39(4):1963–2006
3. Bücher A, Dette H, Volgushev S (2012) A test for Archimedeanity in bivariate copula models. J Multivar Anal 110:121–132
4. Di Lascio FML, Durante F, Jaworski P (2016) Truncation invariant copulas and a testing procedure. J Stat Comput Simul 86(12):2362–2378. doi:10.1080/00949655.2015.1110820
5. Durante F, Jaworski P (2012) Invariant dependence structure under univariate truncation. Statistics 46(2):263–277
6. Durante F, Jaworski P, Mesiar R (2011) Invariant dependence structures and Archimedean copulas. Stat Probab Lett 81(12):1995–2003

7. Fermanian JD (2013) An overview of the goodness-of-fit test problem for copulas. In: Jaworski P, Durante F, Härdle W (eds) Copulae in mathematical and quantitative finance. Lecture notes in statistics. Springer, Berlin, Heidelberg, pp 61–89
8. Genest C, Rémillard B, Beaudoin D (2009) Goodness-of-fit tests for copulas: a review and a power study. Insur Math Econ 44(2):199–213
9. Genest C, Kojadinovic I, Nešlehová J, Yan J (2011) A goodness-of-fit test for bivariate extreme-value copulas. Bernoulli 17(1):253–275
10. Ghoudi K, Rémillard B (2004) Empirical processes based on pseudo-observations. II. The multivariate case. In: Horváth L, Szyszkowicz B (eds) Asymptotic methods in stochastics, vol 44. Fields Institute Communications, American Mathematical Society, Providence, RI, pp 381–406. Proceedings of the international conference (ICAMS'02) held at Carleton University, Ottawa, ON, 23–25 May 2002
11. Gijbels I, Sznajder D (2013) Testing tail monotonicity by constrained copula estimation. Insur Math Econ 52(2):338–351
12. Gijbels I, Omelka M, Sznajder D (2010) Positive quadrant dependence tests for copulas. Can J Stat 38(4):555–581
13. Jaworski P (2010) Testing archimedeanity. In: Borgelt C, González-Rogríguez G, Trutschnig W, Lubiano M, Gil M, Grzegorzewski P, Hryniewicz O (eds) Combining soft computing and statistical methods in data analysis. Advances in intelligent and soft computing, vol 77. Springer, Berlin, pp 353–360
14. Jaworski P (2013) Invariant dependence structure under univariate truncation: the high-dimensional case. Statistics 47(5):1064–1074
15. Jaworski P (2013) The limiting properties of copulas under univariate conditioning. In: Jaworski P, Durante F, Härdle WK (eds) Copulae in mathematical and quantitative finance. Lecture notes in statistics. Springer, Berlin, Heidelberg, pp 129–163
16. Kojadinovic I, Segers J, Yan J (2011) Large-sample tests of extreme-value dependence for multivariate copulas. Can J Stat 39(4):703–720
17. van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes. Springer series in statistics. Springer, New York

# Finite Mixture of Linear Regression Models: An Adaptive Constrained Approach to Maximum Likelihood Estimation

**Roberto Di Mari, Roberto Rocci and Stefano Antonio Gattone**

**Abstract** In order to overcome the problems due to the unboundedness of the likelihood, constrained approaches to maximum likelihood estimation in the context of finite mixtures of univariate and multivariate normals have been presented in the literature. One main drawback is that they require a knowledge of the variance and covariance structure. We propose a fully data-driven constrained method for estimation of mixtures of linear regression models. The method does not require any prior knowledge of the variance structure, it is invariant under change of scale in the data and it is easy and ready to implement in standard routines.

## 1 Introduction

Let $(y_i, \boldsymbol{x}_i)$ be a pair where, respectively, $y_i$ is the random variable of interest and $\boldsymbol{x}_i$ is a vector of $K$ explanatory variables. Finite mixtures of conditional normal distributions can be used to estimate clusterwise regression parameters in a maximum likelihood context. In the literature, clusterwise linear regression is also known under the names of finite mixture of linear regression model or switching regression model [7, 20, 23, 24].

Let the conditional distribution of $y_i|\boldsymbol{x}_i$ be a finite mixture of linear regression models, that is

R. Di Mari (✉) · R. Rocci
DEF, University of Tor Vergata, Rome, Italy
e-mail: roberto.di.mari@uniroma2.it

R. Rocci
e-mail: roberto.rocci@uniroma2.it

S.A. Gattone
DiSFPEQ, University G. d'Annunzio, Chieti-Pescara, Italy
e-mail: gattone@unich.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_23

$$f(y_i|\boldsymbol{x}_i) = \sum_{g=1}^{G} p_g f_g(y_i|\boldsymbol{x}_i, \sigma_g^2, \boldsymbol{\beta}_g) = \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left( - \frac{(y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right), \tag{1}$$

where

(i) $G$ is the number of clusters;
(ii) $\boldsymbol{\beta}_g$ is the vector of regression coefficients for the $g$-th cluster;
(iii) $\sigma_g^2$ is the variance term for the $g$-th cluster.

In addition let us denote the set of parameters to be estimated $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, where $\boldsymbol{\psi} = \{(p_1, \ldots, p_G; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G; \sigma_1^2, \ldots, \sigma_G^2) \in \mathbb{R}^{G(K+2)} : p_1 + \cdots + p_G = 1$, $p_g \geq 0, \sigma_g^2 > 0$ for g $= 1, \ldots, G\}$. Unlike finite mixtures of other densities, the parameters of finite mixtures of linear regression models, under mild regularity conditions [15] are identified.

Let $y_1, \ldots, y_n$, be a sample of independent observations, each respectively observed alongside with a vector of regressors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The likelihood function can be formulated as

$$\mathrm{L}(\boldsymbol{\psi}) = \prod_{i=1}^{n} \left[ \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left( - \frac{(y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right) \right]. \tag{2}$$

Yet, Maximum Likelihood (ML) estimation is known to be problematic: whenever a sample point coincides with the group's center—i.e. its mean—and the group conditional variance approaches zero, the likelihood function increases without bound [5, 21]. Hence a global maximum cannot be found, and the EM algorithm tends to produce the so-called degenerate solutions. Interestingly however, constraining the mixture components to having a unique common variance, although being possibly too restrictive, prevents the likelihood to degenerate.

We propose a constrained solution to the problem of degeneracy, which generalizes the works of Ingrassia [18] and Ingrassia and Rocci [19], and apply it in the context of clusterwise linear regression. We devise an estimation algorithm with data-driven constraints, which are invariant under change of scale in the data and are easy to implement within standard routines. In Sect. 2 we review the literature on the topic of degeneracy, and some of the existing solutions. Section 3 concludes, describing the proposed estimation method.

## 2 The Issue of Degeneracy and How to Avoid It

The likelihood principle is based on the fact that the likelihood function embeds the full information on the parameters contained in the sample. The unboundedness of $L(\psi)$ seems to cause a failure of the maximum likelihood principle. However Kiefer [20] showed that, for switching regressions with components allowed to have component (cluster) specific variances, there is a sequence of estimators which is consistent, asymptotically efficient and normally distributed. This corresponds to a local maximizer in the interior of the parameter space. Yet, even if a local maximum yielding a consistent estimator does exist, there can be several other local maxima. Day [5] showed that, in mixtures with cluster-specific variances, each sample point can generate a singularity in the likelihood function. Similarly, any pair of sample points being sufficiently close together can generate a local maximum—as will triples, quadruplets, etc., which are sufficiently close. This gives rise to a number of spurious maximizers [22], i.e. maximizers which are not good estimates. In the multivariate case, as noticed by Ritter [25], spurious solutions arise from data points being almost coplanar.

Problems related to degeneracy have been tackled by a large number of authors. Possible remedies have conveyed into three main strands, (1) selecting the roots obtained by standard maximum likelihood, or (2) transforming or (3) constraining the likelihood function.

Concerning the first strand, Biernacki and Chrétien [2] provide a domain of attraction leading the estimating algorithm to degeneracy. They show that the speed at which the algorithm converges to an infinite likelihood is at least exponential. As a practical advice, they suggest to run the EM algorithm from different random starts. Biernacki [1] proposes an asymptotic upper bound for the likelihood. Such bound incorporates information on both the sample size and the components variances. Seo and Kim [27] propose to run the EM from several random starts and, at each local maximizer, evaluate the log-likelihood by taking out the $k$ observations with the highest log-likelihood. The root to be selected is the one with the highest $k$-deleted log-likelihood. In a similar fashion, Ritter [25] proposes a method based on Gallegos and Ritter [9], where scale balances are plotted against the likelihood value and a method to select a valid solution among the several ones available is formulated. Such proposal has most natural application in clustering in presence of outliers. Further trimming-based methods can be found in the literature of robust model-based clustering (e.g. [10–12]).

As for the second strand, Chen et al. [3], Ciuperca et al. [4], Eggermont and LaRiccia [8], Green [13], Snoussi and Mohammad-Djafari [26], among the others, address the issue of degeneracy by putting a penalty on the component variances and maximizing the penalized log-likelihood. From a Bayesian perspective, this amounts to incorporating a prior density for the component variances—typically Gamma, for the univariate case, or Wishart, for the multivariate case. Ciuperca et al. [4] prove existence and consistency of the estimator obtained via such penalized maximization.

As for the third strand, Hathaway [14] proposed relative constraints on the variances of the kind

$$\min_{i \neq j} \frac{\sigma_i^2}{\sigma_j^2} \geq c \quad \text{with} \quad c \in (0, 1]. \tag{3}$$

Hathaway's formulation of the maximum likelihood problem presents a strongly consistent global solution, no singularities, a smaller number of spurious maxima. Consistency and robustness of estimators of this sort was already pointed out by

Huber [16, 17]. However, Hathaways constraints are very difficult to apply within iterative procedures like the EM algorithm [6]. To solve this problem, Ingrassia [18] formulated a sufficient condition such that Hathaway's constraints hold, which is easily implementable within the EM algorithm. He shows that constraints in (3) are satisfied when it results

$$a \leq \sigma_g^2 \leq b, \quad \text{with} \quad g = 1, \ldots, G, \tag{4}$$

where $a$ and $b$ are positive numbers such that $a/b \geq c$. In this spirit Ingrassia and Rocci [19] showed how Ingrassia [18] constraints can be implemented directly at each iteration of the EM algorithm, preserving the monotonicity of the algorithm.

The constant $c$ measures the scale balance. As pointed out by Ritter [25], a large scale balance does not deviate dramatically from unique common variance. This in turn means that there is some unknown transformation of the sample space that transfers the component not too far from the common variance setting. High scale variance is indeed valuable, nevertheless it has to be traded with fit.

Concerning the choice of the constant $c$ in (3), among the others, Tan et al. [28], and Xu et al. [29], establish the asymptotic properties of the constrained estimator under a choice of $c$ approaching zero as the sample size increases. Unfortunately, a finite-sample choice of the constant $c$ remains an open issue in all cited works. In addition, as the scale of the data changes, the chosen constant might no longer be appropriate.

In the present work, in the spirit of Hathaway [14] and of the series of papers Ingrassia [18] and Ingrassia and Rocci [19], we formulate constraints where the choice of the constant $c$ is data driven. We exploit the unique common variance by shrinking the component variances towards it, at a shrinkage rate equal to $c$. The resulting constraints are showed to imply Hathaway's. In addition the limitation of lack of invariance of the constraints of Ingrassia and Rocci [19] under change of scale in the response variable in overcome.

## 3 The Proposed Methodology

Starting form the set of constraints of Eq. (4), let $c \in (0, 1]$ and let $\bar{\sigma}^2$ be the unique common variance. The set of constraints proposed in this paper is as follows

$$\sqrt{c} \leq \frac{\sigma_g^2}{\bar{\sigma}^2} \leq \frac{1}{\sqrt{c}},$$

or equivalently

$$\bar{\sigma}^2\sqrt{c} \leq \sigma_g^2 \leq \bar{\sigma}^2\frac{1}{\sqrt{c}}. \tag{5}$$

It is easy to show that (5) implies (3), whereas the converse is not necessarily true, since (5) is more stringent than (3). That is

$$\frac{\sigma_g^2}{\sigma_j^2} = \frac{\sigma_g^2/\bar{\sigma}^2}{\sigma_j^2/\bar{\sigma}^2} \geq \frac{\sqrt{c}}{1/\sqrt{c}} = c.$$

The above constraints still require a choice for the scale balance. Notice that selecting $c$ via ML together with the mixture parameters would trivially yield a scale balance far too optimistic and close to zero—since such a choice would allow for arbitrarily small variances, thus letting the likelihood, or equivalently its log, approach infinity.

We propose to select $c$ using cross-validation, implemented within the estimation routine in Ingrassia and Rocci [19], adapted for clusterwise linear regression. The procedure consists in maximizing with respect to $c$ the cross-validated likelihood. For a given $c$, this is computed as follows.

- Obtain a temporary estimate for the model parameters using the entire sample, which is used as starting value for the cross-validation procedure.
- Partition the full data set into a test set and a training set.
- Estimate the parameters on the training set. Compute the contribution to the log-likelihood of the test set.
- Run the previous two steps $K$ times and sum the contributions of the test sets to the log likelihood.

Such a method, using the constraints of Eq. (5), can be showed to be equivariant under linear affine transformations of the response variable. This is a key property which guarantees that, if the data are linearly transformed, the MLE is transformed accordingly, and the posterior estimates do not change.

In order to assess its validity, the procedure has been tested with an extensive simulation study and an empirical example. The results have shown that our constrained EM algorithm improves upon the unconstrained one and the standard unique common variance, both in terms of accuracy of parameter estimation and clustering.

## References

1. Biernacki C (2004) An asymptotic upper bound of the likelihood to prevent Gaussian mixtures from degenerating. Technical Report, Université de Franche-Comté
2. Biernacki C, Chrétien S (2003) Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with the EM. Stat Probab Lett 61:373–382
3. Chen J, Tan X, Zhang R (2008) Inference for normal mixtures in mean and variance. Statistica Sinica 18(2):443
4. Ciuperca G, Ridolfi A, Idier J (2003) Penalized maximum likelihood estimator for normal mixtures. Scand J Stat 30(1):45–59
5. Day NE (1969) Estimating the components of a mixture of two normal distributions. Biometrika 56:463–474
6. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc: Ser B (Stat Methodol) 39:1–38

7. DeSarbo WS, Cron WL (1988) A maximum likelihood methodology for clusterwise linear regression. J Classif 5(2):249–282
8. Eggermont PPB, LaRiccia VN (2001) Maximum penalized likelihood estimation, vol 1. Springer, New York
9. Gallegos MT, Ritter G (2009) Trimmed ML estimation of contaminated mixtures. Sankhya: Indian J Stat Ser A (2008-):164–220
10. García-Escudero LA, Gordaliza A, Matran C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. Ann Stat 36:1324–1345
11. García-Escudero LA, Gordaliza A, San Martń R, Van Aelst S, Zamar R (2009) Robust linear clustering. J R Stat Soc: Ser B (Stat Methodol) 71(1):301–318
12. García-Escudero LA, Gordaliza A, Mayo-Iscar A, San Martń R (2010) Robust clusterwise linear regression through trimming. Comput Stat Data Anal 54(12):3057–3069
13. Green PJ (1990) On use of the EM for penalized likelihood estimation. J R Stat Soc: Ser B (Stat Methodol) 443–452
14. Hathaway RJ (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Stat 13:795–800
15. Hennig C (2000) Identifiablity of models for clusterwise linear regression. J Classif 17(2):273–296
16. Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1, no 1, pp 221–233
17. Huber PJ (1981) Robust statistics. Wiley, New York
18. Ingrassia S (2004) A likelihood-based constrained algorithm for multivariate normal mixture models. Stat Methods Appl 13:151–166
19. Ingrassia S, Rocci R (2007) A constrained monotone EM algorithm for finite mixture of multivariate Gaussians. Comput Stat Data Anal 51:5339–5351
20. Kiefer NM (1978) Discrete parameter variation: efficient estimation of a switching regression model. Econometrica 46:427–434
21. Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Stat 27:886906
22. McLachlan GJ, Peel D (2000) Finite mixture models. John Wiley and Sons, New York
23. Quandt RE (1972) A new approach to estimating switching regressions. J Am Stat Assoc 67(338):306–310
24. Quandt RE, Ramsey JB (1978) Estimating mixtures of normal distributions and switching regressions. J Am Stat Assoc 73(364):730–738
25. Ritter G (2014) Robust cluster analysis and variable selection. Monographs on statistics and applied probability, vol 137. CRC Press
26. Snoussi H, Mohammad-Djafari A (2001) Penalized maximum likelihood for multivariate Gaussian mixture. In: Fry RL (ed) MaxEnt workshops: Bayesian inference and maximum entropy methods. pp 36–46, Aug 2001
27. Seo B, Kim D (2012) Root selection in normal mixture models. Comput Stat Data Anal 56:2454–2470
28. Tan X, Chen J, Zhang R (2007) Consistency of the constrained maximum likelihood estimator in finite normal mixture models. In: Proceedings of the American Statistical Association, American Statistical Association, Alexandria, VA, 2007, pp 2113–2119 [CD-ROM]
29. Xu J, Tan X, Zhang R (2010) A note on Phillips (1991): "A constrained maximum likelihood approach to estimating switching regressions". J Econom 154:35–41

# A Multivariate Analysis of Tourists' Spending Behaviour

**Marta Disegna, Fabrizio Durante and Enrico Foscolo**

**Abstract** According to the micro-economic theories regarding consumption behaviour, the determinants affecting the joint propensity of purchasing different goods and services are investigated. For this purpose, a copula-based model is suggested to understand how different expenditure categories are dependent with each other. A real application drawn from the tourism field illustrates the proposed approach and shows its main advantages. The findings could guide local practitioners and managers in creating new promotional campaigns able to attract visitors willing to pay on a bundle of goods and services correlated with each other.

## 1 Introduction

The economic impact of tourism flows is often essential for those regions/local communities in which tourism is considered the major source of income [4]. In order to improve the economic effects of tourism visits, appropriate data and tools are needed to study the determinants of tourism expenditure and to analyse the tourists' spending behaviour in depth. In fact, as stated by [1], the use of micro-level makes it possible to observe individual choices regarding the consumption of a tourism commodity or service, and to analyse the heterogeneity and diversity that characterize individual tourism consumption behaviour. In other words, adopting a micro-level approach enables us to take both the consumer behaviour theory on the decision-making process to purchase, and the neoclassical economic theory of budget constraint, into consideration. In particular, we assume that the individual purchase process for a tourism good or service is a two-decision process [5], i.e. the decision to purchase a good

M. Disegna (✉)
Faculty of Management, Bournemouth University, Bournemouth, UK
e-mail: disegnam@bournemmouth.ac.uk

F. Durante · E. Foscolo
Faculty of Economics and Management, Free University of Bozen-Bolzano, Bolzano, Italy
e-mail: fabrizio.durante@unibz.it

E. Foscolo
e-mail: enrico.foscolo@unibz.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_24

followed by the decision on how much to spend on it. The economic theory of budget constraint is based on the assumption of weak separability between goods and services that leads tourists to allocate their budgets in accordance with a three-stage tourist spending process [6]: firstly, tourists decide how much of their budget to allocate for travel; secondly, they decide where to go on vacation; thirdly, they choose how to allocate their tourist budget among various goods and services offered by the selected destination. Obviously, the above-mentioned two economic theories are not disjointed but overlap; this means that an individual has to make a two-stage decision process in each stage comprised in the three-stage tourist spending process. This study contributes to this micro-economic tourism literature by analysing the first stage of the decision-making process (i.e. the propensity of tourist purchase) and the third stage of the budget allocation process (i.e. the allocation of tourist budget among various goods and services offered by a destination) simultaneously. In particular, this paper aims to analyse the factors involved in the decision to consume different categories of tourism goods and services simultaneously. To this end, we exploit the advantages of the copula-based models. Firstly, univariate Logit regressions are estimated per each category by considering a set of possible determinants. Then these regressions are grouped together by means of a copula, which is a multivariate distribution function that aims at describing the dependence among random outcomes in a flexible way [2]. The obtained model allows us to understand whether and how the different purchase decisions are correlated with each other. The methodology is illustrated by analysing a sample of international visitors to the South Tyrol region (Northern Italy).

## 2 The Dataset

The dataset used in this study is drawn from the "International Tourism in Italy" annual survey, conducted by Bank of Italy in order to determine the tourism balance of payments. The survey offers detailed information on the amount of money (in Euro) spent in the five main categories of a typical travel budget: (1) Accommodation ($Y_1$), which also includes expenditure on food and beverages within accommodation premises; (2) Food & beverages ($Y_2$) consumed outside accommodation premises; (3) Shopping ($Y_3$), including souvenirs, gifts, clothes, etc. purchased only for personal use; (4) Internal transportation ($Y_4$) within the visited destination, including purchase of fuel; (5) Other services ($Y_5$), like museums, shows, entertainment, etc. In this study we focus on a subsample of 550 international visitors who spent time in the South Tyrol region (Northern Italy) in 2011 and whose main purpose for the trip was "tourism, holiday, leisure".

The sample consists of 87 % tourists (i.e. people who stayed at least one night in South Tyrol), and 13 % day-visitors. Most of the respondent stated they had incurred costs for tickets and/or transportation fuel (77 %), souvenirs, gifts, items of clothing, or other things for personal use (69 %), and food and beverages (87 %) during the trip to South Tyrol. By contrast, only 34 % of the sample stated they had incurred costs for other services, like museums, shows, entertainments, guided excursions, rented vehicles, or language courses. Table 1 describes the set of explanatory variables ($\mathbf{x}$) considered in this study.

## 3 The Methodology

Let $Y_j$ be a dichotomous variable describing the decision to spend ($Y_j = 1$) or not ($Y_j = 0$) in the $j$-th tourism expenditure category, such as accommodation, transportation, and shopping. This study aims at modelling the dependence among these variables in order to understand whether the decision to spend in one category is correlated with the decision to spend in other categories, given a set of explanatory variables $\mathbf{x}$. Thus, our main interest is to estimate the probability of spending in some (or all) categories given the set of covariates related to the tourist, namely

$$\mathbb{P}\left(Y_1 \leq y_1, \ldots, Y_d \leq y_d \mid \mathbf{x}\right)$$

for $y_j \in \{0, 1\}$. To this end, based on the copula approach, we may assume the relation

$$\mathbb{P}\left(Y_1 \leq y_1, \ldots, Y_d \leq y_d \mid \mathbf{x}\right) = C\left[F_1\left(y_1 \mid \mathbf{x}\right), \ldots, F_d\left(y_d \mid \mathbf{x}\right)\right], \tag{1}$$

where $F_1\left(\cdot \mid \mathbf{x}\right), \ldots, F_d\left(\cdot \mid \mathbf{x}\right)$ are suitable univariate model for, respectively, $Y_1 \mid \mathbf{x}, \ldots, Y_d \mid \mathbf{x}$ (e.g. Logit or Probit model), and $C \in \{C_\theta\}_\theta$ is a suitable copula belonging to a family indexed by the parameter $\theta \in \mathbb{R}^d$. An advantage of this copula model is that the estimation of model parameters for the copula and the regressions can be made in two steps.

1. Univariate models for each marginal distribution are fitted separately. In particular, each $Y_j$ can be described by the logistic regression model specified as follows:

   $$\mathbb{P}(Y_j = 1 \mid \mathbf{x}) = \frac{\exp\left(\mathbf{x}^\top \beta_j\right)}{1 + \exp\left(\mathbf{x}^\top \beta_j\right)} \tag{2}$$

   where $\beta_j$ is the $(j+1)-$dimensional vector including the intercept and the regression coefficients for the $j-$th variable. The estimation of the marginal models is performed by maximum likelihood and the estimates $\widehat{\beta_1}, \ldots, \widehat{\beta_d}$ are obtained. Notice that another binary model (e.g., Probit) may be used as well.
2. A suitable copula $C$ is fitted from a parametric family. Specifically, we suppose that $C$ belongs to the family of multivariate Student $t$–copulas (including the Gaussian copulas as limit cases). These copulas are characterized by a parameter $\nu > 0$, called degree of freedom, and by the parameters of a correlation matrix $(\rho_{k\ell})$, $k, \ell = 1, \ldots, d$. Now, supposing that $\nu$ is held fixed, the estimation of the correlation parameters can be made by solving a system of $d(d-1)/2$ equations related to the score functions (for more details, see [3]).

The copula model considered in the previous steps is quite convenient since the adopted copula family enables us to describe both negative and positive pairwise association among the random variables under consideration. This is particularly useful in our context, since two purchase decisions may be both positively and negatively correlated.

## 4 Model Results and Discussion

Following the two-step estimation method described in Sect. 3, five univariate logistic regression models for each variable in $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)^\top$ were estimated using White's robust standard variance-covariance matrix [7]. The regression models were first estimated considering the whole set of explanatory variables presented in Table 1, then a stepwise procedure was adopted in order to sequentially drop the variables that were non-significant or significant only in one out of five equations for a significance level equal to $\alpha = 0.1$. Table 2 shows the reduced models obtained after this backward procedure.

Once the univariate marginal had been fitted, a score test of independence was performed to check whether the expenditures are conditionally uncorrelated given the explanatory variables. Table 3 shows the corresponding test statistics, denoted by $z_{obs}$. The resulting procedure should reject the assumption of conditional independence if $z_{obs}$ is larger in absolute value than a critical value derived from the standard Gaussian distribution. As can be seen, at a confidence level of 95 %, we can reject the assumption of independence between Accommodation (Y1) and Transportation (Y4), and Accommodation (Y1) and Other services (Y5); while at a confidence level of 90 %, we can reject the assumption of independence between Food & Beverages (Y2) and Other services (Y5), and Transportation (Y4) and Other services (Y5). The other pairs, instead, seem to exhibit a weaker dependence.

Given the values of these statistics, for the sake of illustration we concentrated our attention on the trivariate model formed by the expenditures related to Accommodation (Y1), Transportation (Y4) and Other services (Y5), which exhibit a stronger evidence of association. Table 4 reports the composite likelihood estimates of the pairwise correlations, along with their standard errors assuming either a Gaussian copula or a Student $t$–copula with degrees of freedom equal to 2, 5 or 10, respectively.

To exploit such a model, Fig. 1 reports the estimated probability of spending in all the three considered categories by varying all the explanatory variables, while the "Average level of satisfaction" is fixed at its average value. Analogously, Fig. 2 shows the estimated probabilities by varying all the explanatory variables except for the "Number of paying travelers", which equals its average value. In both cases, the chosen copula model is the Gaussian copula. The models with a Student $t$–copula have also been considered, but the results are similar and, hence, are not reported here.

Regarding the country of origin, visitors from other foreign countries, excluding Germany and Austria, present higher estimated probabilities of spending on the

**Table 1** Description of the explanatory variables

| Independent variable | Description | Mean (Median) |
|---|---|---|
| *Characteristics of the trip* | | |
| Average level of satisfaction | Average level of satisfaction with some aspects of the destination (values from 6 to 10) | 8.334 (8.4) |
| Visit alone | 1 = the respondent makes the trip alone; 0 = otrw | 0.116 (0) |
| Number of paying travellers | Number of travellers who have shared the expenditure of the trip (discrete value from 1 to 7) | 1.945 (2) |
| *Number of times in Italy before* | | |
| Zero | 1 = the interviewee visits any city in Italy for the first time; 0 = otrw | 0.051 (0) |
| Up to 5 times | 1 = been in Italy from 1 to 5 times before the interview; 0 = otrw | 0.229 (0) |
| More than 5 times | 1 = been in Italy more than 5 times before the interview; 0 = otrw (reference category) | 0.720 (1) |
| *Characteristics of the visitor* | | |
| *Country of origin* | | |
| Austrian | 1 = the respondent comes from Austria; 0 = otrw | 0.149 (0) |
| German | 1 = the respondent comes from Germany; 0 = otrw | 0.618 (0) |
| Other country | 1 = the respondent comes from a foreign country; 0 = otrw (reference category) | 0.233 (0) |
| *Employment status* | | |
| Self-employed | 1 = self-employed; 0 = otrw | 0.191 (0) |
| Office worker | 1 = office worker; 0 = otrw | 0.225 (0) |
| Employee | 1 = office employee; 0 = otrw | 0.311 (0) |
| Retired | 1 = retired person; 0 = otrw | 0.220 (0) |
| Other | 1 = other occupation; 0 = otrw (reference category) | 0.054 (0) |
| *Age* | | |
| Less than 35 years old | 1 = less than 35 years old; 0 = otrw (reference category) | 0.131 (0) |
| 35–44 years old | 1 = 35–44 years old; 0 = otrw | 0.267 (0) |
| 45–64 years old | 1 = 45–64 years old; 0 = otrw | 0.425 (0) |
| More than 64 years old | 1 = 65 years old and over; 0 = otrw | 0.176 (0) |

*Notes* For the dichotomous variables, the mean value is to be intended as the proportion of 1's in the sample

**Table 2** Stepwise Logit regression coefficients

| Independent variables | $Y_1^a$ | $Y_2^b$ | $Y_3^c$ | $Y_4^d$ | $Y_5^e$ |
|---|---|---|---|---|---|
| Average level of satisfaction | 1.222***<br>(0.205) | 0.152<br>(0.179) | 0.285**<br>(0.122) | 1.134***<br>(0.174) | 0.005<br>(0.121) |
| Retired | −1.214**<br>(0.358) | −0.337<br>(0.293) | 0.250<br>(0.244) | −0.523**<br>(0.254) | −0.267<br>(0.249) |
| Austrian | −2.294***<br>(0.500) | −1.395**<br>(0.480) | −0.300<br>(0.301) | −1.904***<br>(0.381) | −1.732***<br>(0.353) |
| German | 0.262<br>(0.485) | −0.765*<br>(0.432) | 0.865***<br>(0.233) | −0.309<br>(0.339) | −1.214***<br>(0.220) |
| Number of paying travelers | 0.658**<br>(0.290) | 0.253*<br>(0.148) | 0.155<br>(0.114) | 1.380***<br>(0.363) | 0.021<br>(0.099) |
| Constant | −7.935***<br>(1.680) | 1.128<br>(1.492) | −2.341**<br>(1.004) | −9.629***<br>(1.435) | 0.291<br>(1.032) |

*Notes* ***Significant at $p \leqslant 0.01$, **Significant at $p \leqslant 0.05$, *Significant at $p \leqslant 0.1$. White's robust standard variance-covariance matrix (White, 1980) has been used to estimate the robust standard errors in brackets.

[a] N = 533; Wald $\chi^2(5) = 93.38$; Prob $> \chi^2 = 0$; Log pseudolikelihood = −118.35; McKelvey and Zavoina's $R^2 = 0.52$

[b] N = 533; Wald $\chi^2(5) = 17.83$; Prob $> \chi^2 = 0$; Log pseudolikelihood = −185.41; McKelvey and Zavoina's $R^2 = 0.10$

[c] N = 533; Wald $\chi^2(5) = 39.99$; Prob $> \chi^2 = 0$; Log pseudolikelihood = −307.86; McKelvey and Zavoina's $R^2 = 0.10$

[d] N = 533; Wald $\chi^2(5) = 95.66$; Prob $> \chi^2 = 0$; Log pseudolikelihood = −187.74; McKelvey and Zavoina's $R^2 = 0.57$

[e] N = 533; Wald $\chi^2(5) = 42.96$; Prob $> \chi^2 = 0$; Log pseudolikelihood = −320.18; McKelvey and Zavoina's $R^2 = 0.10$

**Table 3** Score test of independence ($z_{obs}$) among all possible pairs of response variables

| | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ |
|---|---|---|---|---|
| $Y_1$ | 1.4990 | 0.6304 | 3.2730 | 3.3255 |
| $Y_2$ | | −0.5398 | −1.1045 | 1.9434 |
| $Y_3$ | | | 0.7583 | −1.6897 |
| $Y_4$ | | | | 1.8465 |

three expenditure categories simultaneously, regardless of the values assumed by the explanatory variables. Austrian and retired tourists present the lowest estimated probabilities of spending simultaneously. This latter finding can probably be explained by geographical proximity that reduces the probability of their staying at least one night in a South-Tyrolean accommodation for holiday purposes.

Figure 1 reveals that the estimated propensity to spend on Accommodation, Transportation, and on Other services simultaneously increases if the number of paying travellers increases, but only up to four, because for a higher number of paying visitors the propensity becomes quite stable. The Austrian tourists, however, show quite different behaviour since the estimated probability assumes not negligible values

**Table 4** Estimates of the pairwise correlations and their standard errors in four meta-elliptical copula models

| Pair | $t_2$ | | $t_5$ | | $t_{10}$ | | Gaussian | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{\rho}$ | SE | $\hat{\rho}$ | SE | $\hat{\rho}$ | SE | $\hat{\rho}$ | SE |
| $Y_1$–$Y_4$ | −0.8672 | 0.0015 | −0.7117 | 0.0010 | −0.6229 | 0.0004 | −0.5145 | 0.0000 |
| $Y_1$–$Y_5$ | −0.8921 | 0.0040 | −0.7621 | 0.0048 | −0.6746 | 0.0053 | −0.5614 | 0.0059 |
| $Y_4$–$Y_5$ | −0.8526 | 0.0050 | −0.7392 | 0.0048 | −0.6822 | 0.0038 | −0.6200 | 0.0013 |

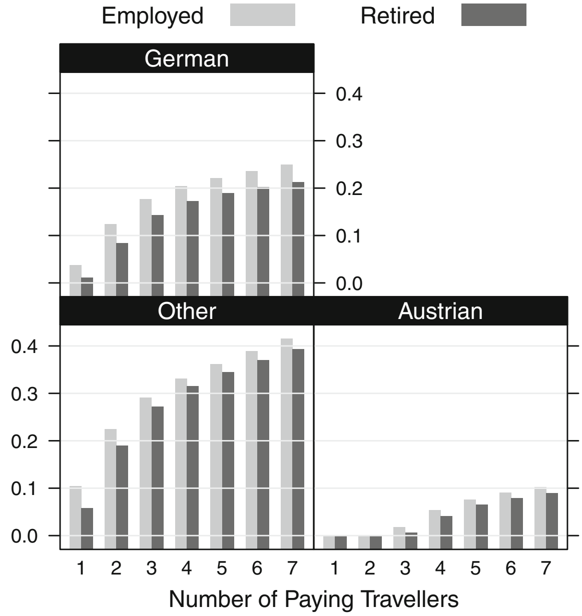


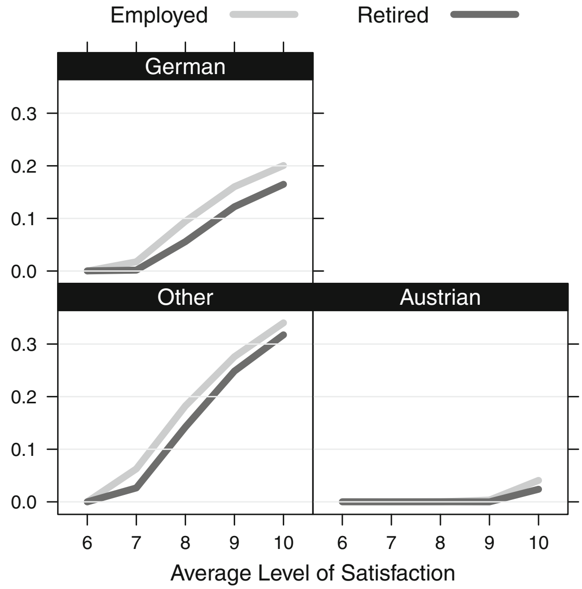
**Fig. 1** Estimated probabilities obtained taken the "Average level of satisfaction" fixed at its average value and varying the other explanatory variables

only when the paying travellers are more than two, until stable levels are reached with bigger groups (i.e. six visitors). This finding is in line with the low estimated probability of spending on the three expenditure categories simultaneously within this group of visitors.

Focusing on Fig. 2 we observe how the estimated probability of spending on Accommodation, Transportation, and Other services simultaneously is significantly affected by the level of satisfaction with the destination. In fact, in the literature it was often recognized that overall satisfaction stimulates higher profitability. The Austrian tourists show an increased estimated probability of spending only for very high satisfaction levels, but, again, this is probably due to the low estimated probability of spending as before.

To summarize, the highest estimated probabilities of spending on the three considered expenditure categories simultaneously was observed for employed foreign tourists (excluding those from Germany or Austria), who are overall very satisfied

**Fig. 2** Estimated probabilities obtained taken the "Number of paying travellers" fixed at its average value and varying the other explanatory variables



with the destination, and who have visited the South-Tyrol in a large groups in which 6 or 7 are paying travellers.

# 5 Conclusions

In this paper a copula–based approach is suggested for studying tourism consumption behaviour, i.e. the probability of spending at a given destination for different goods and services. A sample of international visitors to the South Tyrol region (Northern Italy) in 2011 was analysed to illustrate the main features of the method. The results suggest that a stronger dependence exists between the consumption of Accommodation, Transportation, and Other services; a weaker dependence exists between Food and Beverages, Other services, and Transportation; while the hypothesis of independence is not rejected ($\alpha = 0.1$) for the other combinations of tourism categories.

Focusing on the triplet of expenditures on Accommodation, Transportation, and Other services, the paper illustrates how a set of explanatory variables affects the joint probability of spending simultaneously on these three categories during the same trip. Somehow surprisingly, age and employment status, apart from being retired, do not significantly affect the joint consumption of these three commodities that is affected, on the other hand, by the number of paying travellers, the country of origin of the visitors, and the level of satisfaction towards the destination. To sum up, employed foreign visitors (excluding those from Germany and Austria), who are overall very satisfied with the destination, and who have visited the South-Tyrol in a large group in which 6–7 are paying travellers, present the highest estimated probabilities of simultaneously spending on Accommodation, Transportation, and Other services. Thus, our results highlight that the probability of spending on different tourism goods and services simultaneously is affected not only by economic variables, but also by other socio-demographic and psychographical variables; the level of satisfaction, in particular, plays an important role.

Overall, the findings are of potential interest in tourism management in order to know how visitors decide to allocate their travel budget among different combinations of tourism expenditure categories. Managing this information is fundamental for policy makers and marketing experts in order to improve the touristic supply and to implement specific marketing campaigns that offer a combination of different services (meals, lodging, shopping, etc.) according to tourists' preferences.

## References

1. Alegre J, Pou L (2004) Microeconomic determinants of the probability of tourism consumption. Tour Econ 10(2):125144
2. Durante F, Sempi C (2015) Principles of copula theory. CRC/Chapman and Hall, Boca Raton, FL
3. Genest C, Nikoloulopoulos A, Rivest L, Fortin M (2013) Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. Braz J Probab Stat 27(3):265284
4. Hung W-T, Shang J-K, Wang F-C (2012) Another look at the determinants of tourism expenditure. Ann Tour Res 39(1):495–498
5. Pudney S (1989) Modelling individual choice: the econometrics of corners, Kinks, and Holes. Basil Blackewll, London
6. Syriopoulos T, Sinclair M (1993) An econometric study of tourism demand: the AIDS model of US and European tourism in Mediterranean countries. Appl Econ 25(12):15411552
7. White H (1980) A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica 48(4):817838

# Robust Fuzzy Clustering via Trimming and Constraints

**Francesco Dotto, Alessio Farcomeni, Luis Angel García-Escudero and Agustín Mayo-Iscar**

**Abstract** A methodology for robust fuzzy clustering is proposed. This methodology can be widely applied in very different statistical problems given that it is based on probability likelihoods. Robustness is achieved by trimming a fixed proportion of “most outlying” observations which are indeed self-determined by the data set at hand. Constraints on the clusters’ scatters are also needed to get mathematically well-defined problems and to avoid the detection of non-interesting spurious clusters. The main lines for computationally feasible algorithms are provided and some simple guidelines about how to choose tuning parameters are briefly outlined. The proposed methodology is illustrated through two applications. The first one is aimed at heterogeneously clustering under multivariate normal assumptions and the second one might be useful in fuzzy clusterwise linear regression problems.

## 1 Introduction

Hard clustering methods are aimed at searching meaningful partitions of a data set into $k$ disjoint clusters. Therefore, “0–1” membership values of observations to clusters are provided. On the other hand, fuzzy clustering methods provide nonnegative membership values which may generate overlapping clusters where every subject is shared among all clusters [2, 28].

It is known that the presence of an (even a small) amount of outlying observations can be problematic when applying traditional hard clustering methods. For instance,

F. Dotto · A. Farcomeni
Sapienza University of Rome, Rome, Italy
e-mail: francesco.dotto@uniroma1.it

A. Farcomeni
e-mail: alessio.farcomeni@uniroma1.it

L.A. García-Escudero (✉) · A. Mayo-Iscar
University of Valladolid, Valladolid, Spain
e-mail: lagarcia@eio.uva.es

A. Mayo-Iscar
e-mail: agustim@eio.uva.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_25

clearly differentiated clusters can be wrongly joined together and non-interesting clusters (made up of only few outlying observations) can be detected. This is also the case when applying many fuzzy clustering techniques. In fact, historically, the fuzzy clustering community was the first one to face this robustness issue. This is due to the fact that outliers may be approximately "equally remote" from all clusters and, thus, they may have similar (but not necessarily small) membership values.

References on robustness in hard clustering can be found in [10] and in two recent [7, 24] books. On the other hand, [1, 4] are good reviews on robust fuzzy clustering. These proposals in fuzzy clustering include "noise clustering" [3], the replacement of the Euclidean distance by other discrepancy measures [22, 31] or the use of "possibilistic" clustering [19].

Trimming has a long history as a simple way to provide robustness to statistical procedures. Its application in clustering needs to be done by taking into account the possibility of discarding "bridge points". A sensible way to perform trimming is to let the data decide which observations must be trimmed such that we find an optimal clustering for the non-trimmed ones. This is the "impartial" trimming approach adopted when using the TCLUST method [9]. This approach was extended in [8] to fuzzy clustering. This can be also seen as an extension of the "least trimmed squares" approach in fuzzy clustering [17]. Discarding a fixed fraction of data was also considered in [18].

One clear advantage of the methodology in [8] is that it allows the detection of non-necessarily spherically-shaped clusters. Additionally, the use of likelihoods in its statement allows its generalization to very different frameworks. The use of procedures based on likelihoods is not new in fuzzy clustering (see, e.g., [12, 13, 25, 26, 30, 32]). Note also that some type of constraint on the clusters' scatters is always needed. Otherwise, the defining problem would become a mathematically ill-posed one. By using these constraints, clusters with arbitrarily very different scatters are not allowed. The use of procedures based on likelihoods is also useful in clusterwise linear regression problems. Instead of detecting clusters just around centroids, it is often interesting to detect clusters around linear structures [15, 21, 29] (hard clustering) and [14, 16] (fuzzy clustering).

## 2 Methodology

Suppose that we have $n$ observations $\{x_1, \dots, x_n\}$ in $\mathbb{R}^p$ and we want to group them into $k$ clusters in a fuzzy way. Therefore, our aim is to obtain a collection of nonnegative membership values $u_{ij} \in [0, 1]$ for all $i = 1, \dots, n$ and $j = 1, \dots, k$. A membership value 1 indicates that object $i$ fully belongs to cluster $j$ while a 0 membership value means that it does not belong at all to this cluster. However, intermediate degrees of membership are allowed when $u_{ij} \in (0, 1)$. We consider that an observation is fully trimmed if $u_{ij} = 0$ for all $j = 1, \dots, k$.

Let us assume that $\varphi(\cdot;\theta_j)$ is a $p$-variate probability density function in $\mathbb{R}^p$ that depends on a set of parameters $\theta_j$. Given a fixed trimming level $\alpha \in [0,1)$ and a fixed value of the fuzzifier parameter $m>1$; a robust constrained fuzzy clustering problem can be defined through the maximization of:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{m}\log(p_j\varphi(x_i;\theta_j)), \tag{1}$$

where the membership values $u_{ij}\geq 0$ are assumed to satisfy

$$\sum_{j=1}^{k} u_{ij}=1 \text{ if } i\in\mathcal{I} \text{ and } \sum_{j=1}^{k} u_{ij}=0 \text{ if } i\notin\mathcal{I},$$

for a subset $\mathcal{I}\subset\{1,2,\ldots,n\}$ with $\#\mathcal{I}=[n(1-\alpha)]$, when $\theta=(\theta_1,\ldots,\theta_k)\in\Theta$, for a given parametric space $\Theta$, and the $p_j$'s are positive weights satisfying $\sum_{j=1}^{k} p_j=1$. Notice that $u_{i1}=\cdots=u_{ik}=0$ for all $i\notin\mathcal{I}$, so these observations do not contribute to the summation in (1). The notation $[\cdot]$ is used for the floor function.

For instance, we may consider $\theta_j=(m_j,S_j)$ and

$$\varphi(x_i;\theta_j)=(2\pi)^{-p/2}|S_j|^{-1}\exp\big(-(x_i-m_j)'S_j^{-1}(x_i-m_j)/2\big). \tag{2}$$

In a clusterwise linear regression framework, if $x_i=(y_i,\mathbf{x}_i')$ with $y_i\in\mathbb{R}$ as the response variable value and $\mathbf{x}_i\in\mathbb{R}^{p-1}$ as the values taken by $p-1$ explanatory variables, then we can use $\theta_j=(\beta_j,s_j^2)$ and

$$\varphi(x_i;\theta_j)=(2\pi s_j^2)^{-1/2}\exp\big(-(y_i-\mathbf{x}_i'\beta_j)^2/(2s_j^2)\big). \tag{3}$$

In the target function (1), clusters' weights $p_j$'s are also included. This may be seen as an "entropy regularization" [23]. Including these weights is interesting when the number of clusters is misspecified, because some $p_j$ weights can be set close to 0 when $k$ is larger than the "true" number of clusters. Another possibility is to exclude these weights by directly assuming $p_1=\cdots=p_k=1/k$. This would shrink assignments towards similar number of observations within each cluster.

It is important to note that the maximization of (1) when $k>1$ is commonly an ill-posed problem without any constraint on the scatter parameters. For instance, in the two previous problems, we can see that (1) becomes unbounded when $|S_j|\to 0$ or when $s_j^2\to 0$. Additionally, these constraints are useful to avoid the detection of non-interesting "spurious" solutions. Thus, in [8], it is proposed the use of an eigenvalue ratio constraint

$$\frac{\max_{j=1}^{k}\max_{l=1}^{p}\lambda_l(S_j)}{\min_{j=1}^{k}\min_{l=1}^{p}\lambda_l(S_j)}\leq c, \tag{4}$$

for a fixed constant $c \geq 1$, where $\{\lambda_l(S)\}_{l=1}^{p}$ denote the $p$ eigenvalues of the matrix $S$. In a similar way, the use of (3) with the constraint

$$\frac{\max_{j=1}^{k} s_j^2}{\min_{j=1}^{k} s_j^2} \leq c, \tag{5}$$

is proposed in [6] for fuzzy clusterwise linear clustering.

Therefore, if $\Theta_c \subseteq \Theta$ denotes the restricted parametric space, the maximization of (1) when $\theta \in \Theta_c$ yields the FTCLUST method ($\varphi(\cdot)$ as in (2) and (4)) and the FTCLUST-R method ($\varphi(\cdot)$ as in (3) and (5)).

## 3 Algorithm

The maximization of (1) under those constraints is not an easy problem. However, a feasible algorithm can be given:

1. *Initialization*: The procedure is initialized several times by randomly selecting initial $\theta_j$'s parameters. This can be done by selecting $k$ subsets of size $p+1$ in general position. Fitting $k$ simple models within each subsample allows to obtain these initial $\theta_j$'s. Weights $p_1, \dots, p_k$ with $p_j \in (0, 1)$ and summing up to 1 are also randomly chosen.
2. *Iterative steps*: The following steps are executed until convergence or a maximum number of iterations is reached.

   2.1. *Membership values*: If $\max_{j=1,\dots,k} p_j\varphi(x_i;\theta_j) \geq 1$, then

   $$u_{ij} = I\big\{p_j\varphi(x_i;\theta_j) = \max_{q=1,\dots,k} p_q\varphi(x_i;\theta_q)\big\} \text{ (hard assignment)},$$

   with $I\{\cdot\}$ as the 0–1 indicator function. If $\max_{q=1,\dots,k} p_q\varphi(x_i;\theta_q) < 1$, then

   $$u_{ij} = \left(\sum_{q=1}^{k}\left(\frac{\log(p_j\varphi(x_i;\theta_j))}{\log(p_q\varphi(x_i;\theta_q))}\right)^{\frac{1}{m-1}}\right)^{-1} \text{ (fuzzy assignment)}.$$

   2.2. *Trimmed observations*: Let

   $$r_i = \sum_{j=1}^{k} u_{ij}^{m} \log(p_j\varphi(x_i;\theta_j)) \tag{6}$$

   and $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$ be these values sorted. The observations to be trimmed are those with indexes $\{i : r_i < r_{([n\alpha])}\}$. The membership values for those observations are redefined as $u_{ij} = 0$, for every $j$ if $r_i < r_{([n\alpha])}$.

2.3. *Update parameters*: Given the membership values obtained in the previous step, the parameters are updated as

$$p_j = \sum_{i=1}^{n} u_{ij}^m \Big/ \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m,$$

and the $\theta_j$'s are updated by maximizing (1) where the $u_{ij}$'s are those obtained in the previous step. For instance, this maximization implies the use of weighted means and weighted covariance matrices for the FTCLUST and the use of weighted least squares for the FTCLUST-R (weights $u_{ij}^m$ in both cases). In more general frameworks, a weighted likelihood should be maximized in a closed form or numerically.
It may happen that these so obtained $\theta_j$'s do not fall within $\Theta_c$. In this case, as done in [8] and [6], it is needed to modify them properly by using optimally truncated scatter parameters. I.e., if $\{d_l\}$ are these scatter parameters (eigenvalues in the case of the FTCLUST and error terms' variances in the case of FTCLUST-R), then we use

$$[d_l]_t = \begin{cases} d_l & \text{if } d_l \in [t, ct] \\ t & \text{if } d_l < t \\ ct & \text{if } d_l > ct \end{cases},$$

with $t$ being a threshold value. Note that these truncated $\{d_l\}$ do satisfy the required constraints and we only need to obtain the optimal threshold value $t_{opt}$ which maximizes (1). Sometimes, there are closed forms expressions for obtaining $t_{opt}$ (see [8] and [6]).

3. *Evaluate objective function* and return parameters yielding the highest (1).

This algorithm can be seen as a fuzzy extension of the classical EM algorithm [5] where "concentration steps", as those in [27], are also applied. Note also that it naturally leads to a fuzzy clustering method with "high contrast" [25] (a compromise between "hard" and "fuzzy" clustering methods).

## 4 Tuning Parameters

The proposed methodology exhibits high flexibility but the price we pay is that of specifying several tuning parameters. In this section, we briefly discus about them and we give some practical guidelines for their choice.

*Fuzzifier parameter*: Parameter $m$ serves to control the degree of fuzziness in the obtained clustering. The $m = 1$ case provides "hard" or "crisp" clustering membership values. In fact, with $m = 1$, we recover the TCLUST method in [9] from the FTCLUST and the robust linear grouping in [11] (without second trimming) from the FTCLUST-R. However, there is an unexpected problem if $m > 1$ when applying fuzzy clustering approaches based on the maximum likelihood principle. This inherent problem has to do with the different effect of $m$ depending on the scale (i.e., when we replace $x_i$ by $S \cdot x_i$ for a given constant $S$). This problem can be addressed by choosing simultaneously $m$ and the scale of data ($S$) in such a way that we achieve some pre-specified "proportions of hard assignments" and "relative entropy". The relative entropy is defined as $\sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \log u_{ij} / [n(1-\alpha)] \log(k)$.

*Trimming level*: The trimming level $\alpha$ is the proportion of observations discarded. Although an $\alpha$ value smaller than the true contamination level can be problematic, we can see that $\alpha$ (slightly) higher than needed most of times provides good $\theta_j$ estimates. Then, wrongly trimmed observations can be recovered back. Additionally, given a tentative $\alpha$ value and $r_{(1)} \leq \cdots \leq r_{(n)}$ being the sorted $r_i$ values in (6), we can check if this $\alpha$ was a sensible choice by seeing whether these $r_{(i)}$ increase quickly when $i/n < \alpha$ and increase slowly when $i/n > \alpha$.

*Constraint on the scatter parameters*: The constant $c$ serves to control the degree of "heteroscedasticity" in the obtained clusters. A large $c$ value allows for more different variances in the error terms when using FTCLUST-R. Large $c$ values also allows for more severe departures from sphericity in FTCLUST. The most constrained case $c = 1$ (with $\alpha = 0$ and "equal weights") yields the classical fuzzy $k$-means [2] when using FTCLUST and fuzzy $k$-regressions [14] when using FTCLUST-R.

## 5 An Example

We conclude with an example of the application of FTCLUST to the "M5data" set in [9] (available at the `tclust` package in the CRAN repository). This data set is obtained from three normal bivariate distributions with different scales and proportions (see the "true" cluster labels in Fig. 1a). One of the components strongly overlaps with another one and there is a 10 % background noise. Figure 1b shows the very bad results obtained when applying FTCLUST with $\alpha = 0$ (all observations are wrongly shared with similar membership values). We can see in Fig. 1c that the use $\alpha = 0.1$ and $c = 1$ gives better clustering results but it is unable to deal with the very different cluster scatters. Finally, Fig. 1d shows the excellent results obtained $\alpha = 0.1$ and $c = 50$, i.e. a higher eigenvalues ratio constraint value.

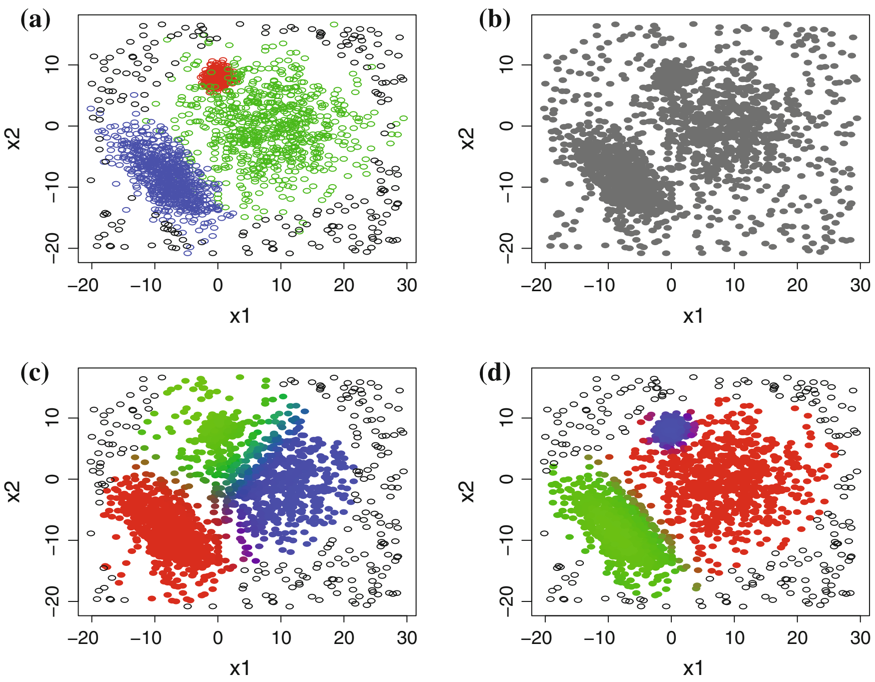


**Fig. 1** **a** “M5data” dataset with the true assignments. Results of applying FTCLUST with $\alpha = 0$ and $c = 1$ in (**b**), $\alpha = 0.1$ and $c = 1$ in (**c**) and $\alpha = 0.1$ and $c = 50$ in (**d**). A mixture of *red*, *blue* and *green* colors with intensities proportional to the membership values are used to summarize the clustering results and “∘” are the trimmed observations

**Acknowledgments** Research partially supported by the Spanish Ministerio de Economía y Competitividad, grant MTM2014-56235-C2-1-P, and by Consejería de Educación de la Junta de Castilla y León, grant VA212U13.

## References

1. Banerjee A, Davé RN (2012) Robust clustering. Wires Data Min Knowl 2:29–59
2. Bezdek JC (1981) Pattern recognition with fuzzy objective function algoritms. Plenum Press, New York
3. Davé RN (1991) Characterization and detection of noise in clustering. Pattern Recogn Lett 12:657–664
4. Davé RN, Krishnapuram R (1997) Robust clustering methods: a unified view. IEEE Trans Fuzzy Syst 5:270–293
5. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38
6. Dotto F, Farcomeni A, García-Escudero LA, Mayo-Iscar A (2016) A fuzzy approach to robust regression clustering. Submitted manuscript

7. Farcomeni A, Greco L (2015) Robust methods for data reduction. Chapman and Hall/CRC, Boca Raton, Florida
8. Fritz H, García-Escudero LA, Mayo-Iscar A (2013) Robust constrained fuzzy clustering. Inf Sci 245:38–52
9. García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. Ann Stat 36:1324–1345
10. García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2010) A review of robust clustering methods. Adv Data Anal Classif 4:89–109
11. García-Escudero LA, Gordaliza A, San Martín R, Mayo-Iscar A (2010) Robust clusterwise linear regresin through trimming. Comput Stat data Anal 54:3057–3069
12. Gath I, Geva AB (1989) Unsupervised optimal fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 11:773–781
13. Gustafson EE, Kessel WC (1979) Fuzzy clustering with a fuzzy covariance matrix. Proceedings of the IEEE lnternational conference on fuzzy systems, San Diego, pp 761–766 (1979)
14. Hathaway RJ, Bezdek JC (1993) Switching regression models and fuzzy clustering. IEEE Trans Fuzzy Syst 1:195–204
15. Hosmer DW (1974) Maximun likelihood estimates of the parameters of a mixture of two regression lines. Commun Stat Theory Methods 3:995–1006
16. Kuo-Lung W, Miin-Shen Y, June-Nan H (2009) Alternative fuzzy switching regression. In: Proceedings of the international multiconference of engineers and computer scientist
17. Kim J, Krishnapuram R, Davé R (1996) Application of the least trimmed squares technique to prototype-based clustering. Pattern Recogn Lett 17:633–641
18. Klawonn F (2004) Noise clustering with a fixed fraction of noise. In: Lotfi A, Garibaldi JM (eds) Applications and science in soft computing. Springer, Berlin-Heidelberg, pp 133–138
19. Krishnapuram R, Keller JM (1993) A possibilistic approach to clustering. IEEE Trans Fuzzy Syst 1:98–110
20. Krishnapuram R, Keller JM (1996) The possibilistic $C$-means algorithm: Insights and recommandations. IEEE Trans Fuzzy Syst 4:385–393
21. Lenstra AK, Lenstra JK, Rinnoy Kan AHG, Wansbeek TJ (1982) Two lines least squares. Ann Discrete Math 16:201–211
22. Łeski J (2003) Towards a robust fuzzy clustering. Fuzzy Set Syst 137:215–233
23. Miyamoto S, Mukaidono M (1997) Fuzzy $c$-means as a regularization and maximum entropy approach. In: Proceedings of the 7th international fuzzy systems association world congress (IFSA'97), pp 86–92
24. Ritter G (2015) Robust cluster analysis and variable selection. Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton, Florida
25. Rousseeuw PJ, Trauwaert E, Kaufman L (1995) Fuzzy clustering with high contrast. J Comput Appl Math 64:81–90
26. Rousseeuw PJ, Kaufman L, Trauwaert E (1996) Fuzzy clustering using scatter matrices. Comput Stat Data Anal 23:135–151
27. Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
28. Ruspini E (1969) A new approach to clustering. Inf Control 15:22–32
29. Späth H (1982) A fast algorithm for clusterwise regression. Computing 29:175–181
30. Trauwaert E, Kaufman L, Rousseeuw PJ (1991) Fuzzy clustering algorithms based on the maximum likelihood principle. Fuzzy Sets Syst 42:213–227
31. Wu KL, Yang MS (2002) Alternative $c$-means clustering algorithms. Pattern Recogn 35:2267–2278
32. Yang MS (1993) On a class of fuzzy classification maximum likelihood procedures. Fuzzy Set Syst 57:365–337

# One-Factor Lévy-Frailty Copulas with Inhomogeneous Trigger Rates

**Janina Engel, Matthias Scherer and Leonhard Spiegelberg**

**Abstract** A new parametric family of high-dimensional, non-exchangeable extreme-value copulas is presented. The construction is based on the Lévy-frailty construction and stems from a subfamily of the Marshall–Olkin distribution. In contrast to the classical Lévy-frailty construction, non-exchangeability is achieved by inhomogeneous trigger-rate parameters. This family is studied with respect to its distributional properties and a sampling algorithm is developed. Moreover, a new estimator for its parameters is given. The estimation strategy consists in minimizing the mean squared error of the underlying Bernstein function and certain strongly consistent estimates thereof.

**Keywords** Extreme-value copula · Non-exchangeable Lévy-frailty model

## 1 Motivation

The Marshall–Olkin distribution, see [7], is a cornerstone in reliability theory and quantitative risk management. It extends the exponential law to higher dimensions by maintaining its lack-of-memory property. Its conditionally i.i.d. subfamily is analyzed in [5] and an alternative construction—termed Lévy-frailty model—is given. This model is generalized in the present manuscript by allowing for inhomogeneous intensity parameters for the exponentially distributed trigger variables. In this way, non-exchangeability is achieved, which carries over to the induced survival copula. Most parametric families of high-dimensional copulas are exchangeable, explaining why research relaxing this assumption is necessary, see, e.g., [4, p.14] for a multi-factor Lévy-frailty construction and [6] for applications.

J. Engel · M. Scherer (✉) · L. Spiegelberg
Technische Universität München, Garching-Hochbrück, Germany
e-mail: scherer@tum.de

J. Engel
e-mail: janina.engel@tum.de

L. Spiegelberg
e-mail: spiegelb@in.tum.de

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_26

An important field of application for such models is portfolio-credit risk. The default of a company can be triggered by two sources: (a) company-individual risk factors—modeled by independent exponentially distributed trigger variables—and (b) market risk factors—represented by a Lévy subordinator that is acting as (common) stochastic clock and thus introducing the underlying dependence structure. In such a situation, non-homogeneous credit spreads can be achieved via individual intensity-rate parameters.

The remainder of this article is organized as follows. Section 2 explains the stochastic model, derives the Lévy-frailty copula (LFC) with inhomogeneous trigger rate parameters, summarizes its statistical properties, and provides a simulation algorithm. Section 3 introduces a new estimator for the parameters of the LFC. Finally, Sect. 4 concludes.

## 2 Generalized One-Factor Lévy-Frailty Copulas

A new family of one-factor Lévy-frailty copulas, with a non-trivial Lévy subordinator $\Lambda = \{\Lambda_t\}_{t\geq 0}$ and inhomogeneous, exponentially distributed trigger variables $\{E_k\}_{k\in\mathbb{N}}$ with rate parameters $\{\lambda_k\}_{k\in\mathbb{N}}$ as stochastic building blocks, is introduced. A standard reference for Lévy subordinators is [1]. While $\Lambda$ serves as a common time-change, the individual trigger variables $\{E_k\}_{k\in\mathbb{N}}$ model the arrival times of specific events. The sequence of random variables $\{X_k\}_{k\in\mathbb{N}}$ is defined as the collection of first-passage times of $\Lambda$ across the thresholds $\{E_k\}_{k\in\mathbb{N}}$, i.e.

$$X_k := \inf\{t > 0 : \Lambda_t > E_k\}, \quad k \in \mathbb{N}. \tag{1}$$

A vivid example for this construction is to interpret $X_k$ as the default time of a risky asset that is triggered by the occurrence of shock $E_k$. So in general $X_k$ can be understood as a future time point at which an event related to asset $k$ takes place. The dependence structure lurking behind this construction is characterized in the following.

**Theorem 1** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which the following independent objects are defined. (a) The list $E_1, \dots, E_d$ of independent, exponential random variables with $E_k \sim \mathcal{E}(\lambda_k)$, (b) $\Lambda = \{\Lambda_t\}_{t\geq 0}$ a Lévy subordinator with Laplace exponent $\Psi_\Lambda$, i.e. $\mathbb{E}[\exp(-x\Lambda_t)] = \exp(-t\Psi_\Lambda(x))$, $x, t \geq 0$, excluding $\Lambda_t \equiv 0$. Further, we define the random variables*

$$X_k := \inf\{t > 0 : \Lambda_t > E_k\}, \quad k = 1, \dots, d.$$

*It follows that $X_k \sim \mathcal{E}(\Psi_\Lambda(\lambda_k))$ is exponentially distributed and the survival copula $\hat{C}$ of $(X_1, \dots, X_d)$ is given by*

$$\hat{C}(u_1, \dots, u_d) = \prod_{i=1}^{d} u_{\pi(i)}^{\frac{\Psi_\Lambda\left(\sum_{j=1}^{i} \lambda_{\pi(j)}\right) - \Psi_\Lambda\left(\sum_{j=1}^{i-1} \lambda_{\pi(j)}\right)}{\Psi_\Lambda\left(\lambda_{\pi(i)}\right)}},$$

*where* $\Psi_\Lambda\big(\sum_{j=1}^{0}\lambda_{\pi(j)}\big)=0$*, i.e. the exponent of* $u_{\pi(1)}$ *equals* $1$*, and* $\pi:\{1,\dots,d\}\to\{1,\dots,d\}$ *is a permutation depending on* $\overrightarrow{u}:=(u_1,\dots,u_d)$*,* $\Psi_\Lambda$*, and* $\lambda_1,\dots,\lambda_d$ *such that*

$$u_{\pi(1)}^{\frac{1}{\Psi_\Lambda(\lambda_{\pi(1)})}} \le u_{\pi(2)}^{\frac{1}{\Psi_\Lambda(\lambda_{\pi(2)})}} \le \cdots \le u_{\pi(d)}^{\frac{1}{\Psi_\Lambda(\lambda_{\pi(d)})}}. \tag{2}$$

*Furthermore, the random vector* $(U_1,\dots,U_d)$ *with joint distribution function* $\hat{C}$ *is given by* $(\exp(-\Psi_\Lambda(\lambda_1)X_1),\dots,\exp(-\Psi_\Lambda(\lambda_d)X_d))$.

*Proof* We start with the marginal laws, showing $X_k\sim\mathcal{E}(\Psi_\Lambda(\lambda_k))$. Let $t>0$

$$\mathbb{P}(X_k>t)=\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{E_k>\Lambda_t\}}\mid\sigma(\Lambda_t)\right]\right]=\mathbb{E}\left[e^{-\lambda_k\Lambda_t}\right]=e^{-t\Psi_\Lambda(\lambda_k)}.$$

The joint survival probability of $(X_1,\dots,X_d)$ can be derived similarly. Let $t_1>0,\dots,t_d>0$ and let $\pi$ be a permutation such that $t_{\pi(d)}\le t_{\pi(d-1)}\le\cdots\le t_{\pi(1)}$ holds. Then, by the tower rule and conditional independence

$$\begin{aligned}
&\mathbb{P}(X_1>t_1,X_2>t_2,\dots,X_d>t_d)\\
&=\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\Lambda_{t_{\pi(d)}}<E_{\pi(d)},\dots,\Lambda_{t_{\pi(2)}}<E_{\pi(2)},\Lambda_{t_{\pi(1)}}<E_{\pi(1)}\}}\mid\sigma(\Lambda)\right]\right]\\
&=\mathbb{E}\left[\mathbb{P}\left(\Lambda_{t_{\pi(d)}}<E_{\pi(d)}\mid\sigma(\Lambda)\right)\cdot\ldots\cdot\mathbb{P}\left(\Lambda_{t_{\pi(1)}}<E_{\pi(1)}\mid\sigma(\Lambda)\right)\right]\\
&=\mathbb{E}\left[\exp\left(-\lambda_{\pi(d)}\Lambda_{t_{\pi(d)}}-\cdots-\lambda_{\pi(1)}\Lambda_{t_{\pi(1)}}\right)\right]
\end{aligned}$$

and we observe that $\Lambda_{t_{\pi(2)}}=\Lambda_{t_{\pi(d)}}+\big(\Lambda_{t_{\pi(d-1)}}-\Lambda_{t_{\pi(d)}}\big)+\cdots+\big(\Lambda_{t_{\pi(2)}}-\Lambda_{t_{\pi(3)}}\big)$ and $\Lambda_{t_{\pi(1)}}=\Lambda_{t_{\pi(d)}}+\big(\Lambda_{t_{\pi(d-1)}}-\Lambda_{t_{\pi(d)}}\big)+\cdots+\big(\Lambda_{t_{\pi(1)}}-\Lambda_{t_{\pi(2)}}\big)$, similarly for the other involved quantities. So we continue with the derivation of the joint survival probability and find

$$\begin{aligned}
\cdots&=\mathbb{E}\left[\exp\left(-(\lambda_{\pi(d)}+\lambda_{\pi(d-1)}+\cdots+\lambda_{\pi(1)})\Lambda_{t_{\pi(d)}}\right)\right]\\
&\quad\cdot\mathbb{E}\left[\exp\left(-(\lambda_{\pi(d-1)}+\cdots+\lambda_{\pi(1)})\left(\Lambda_{t_{\pi(d-1)}}-\Lambda_{t_{\pi(d)}}\right)\right)\right]\cdot\ldots\\
&\quad\cdot\mathbb{E}\left[\exp\left(-(\lambda_{\pi(1)})\left(\Lambda_{t_{\pi(1)}}-\Lambda_{t_{\pi(2)}}\right)\right)\right]\\
&=\mathbb{E}\left[\exp\left(-(\lambda_{\pi(d)}+\lambda_{\pi(d-1)}+\cdots+\lambda_{\pi(1)})\Lambda_{t_{\pi(d)}}\right)\right]\\
&\quad\cdot\mathbb{E}\left[\exp\left(-(\lambda_{\pi(d-1)}+\cdots+\lambda_{\pi(1)})\left(\Lambda_{t_{\pi(d-1)}-t_{\pi(d)}}\right)\right)\right]\cdot\ldots\\
&\quad\cdot\mathbb{E}\left[\exp\left(-(\lambda_{\pi(1)})\left(\Lambda_{t_{\pi(1)}-t_{\pi(2)}}\right)\right)\right]
\end{aligned}$$

$$\begin{aligned}
&=\exp\Big(-t_{\pi(d)}\Psi_\Lambda(\lambda_{\pi(d)}+\lambda_{\pi(d-1)}+\cdots+\lambda_{\pi(1)})\Big)\\
&\quad\cdot\exp\Big(-(t_{\pi(d-1)}-t_{\pi(d)})\Psi_\Lambda(\lambda_{\pi(d-1)}+\cdots+\lambda_{\pi(1)})\Big)\cdot\ldots\\
&\quad\cdot\exp\Big(-(t_{\pi(1)}-t_{\pi(2)})\Psi_\Lambda(\lambda_{\pi(1)})\Big)\\
&=\prod_{i=1}^{d}\exp\Big(-t_{\pi(i)}\Big[\Psi_\Lambda\big(\sum_{j=1}^{i}\lambda_{\pi(j)}\big)-\Psi_\Lambda\big(\sum_{j=1}^{i-1}\lambda_{\pi(j)}\big)\Big]\Big).
\end{aligned}$$

**Table 1** Properties and dependence measures of the LFC $\hat{C}$

| | |
|---|---|
| Stable tail dependence function | $l(x_1,\ldots,x_d) := \sum_{i=1}^{d} x_{\pi(i)} \frac{\Psi_\Lambda\left(\sum_{j=1}^{i}\lambda_{\pi(j)}\right)-\Psi_\Lambda\left(\sum_{j=1}^{i-1}\lambda_{\pi(j)}\right)}{\Psi_\Lambda(\lambda_{\pi(i)})}$ |
| Pickands representation ($d = 2$) | $A(x) = \begin{cases} 1+\frac{x\psi}{\Psi_\Lambda(\lambda_2)} & \text{if } x \leq \frac{\Psi_\Lambda(\lambda_2)}{(\Psi_\Lambda(\lambda_1)+\Psi_\Lambda(\lambda_2))} \\ 1+\frac{\psi(1-x)}{\Psi_\Lambda(\lambda_1)} & \text{if } x > \frac{\Psi_\Lambda(\lambda_2)}{(\Psi_\Lambda(\lambda_1)+\Psi_\Lambda(\lambda_2))} \end{cases}$, where<br>$\psi = \Psi_\Lambda(\lambda_1+\lambda_2) - \Psi_\Lambda(\lambda_1) - \Psi_\Lambda(\lambda_2)$ |
| Spearman's $\rho$ | $\rho_S = 3 \cdot \frac{\Psi_\Lambda(\lambda_1)+\Psi_\Lambda(\lambda_2)-\Psi_\Lambda(\lambda_1+\lambda_2)}{\Psi_\Lambda(\lambda_1+\lambda_2)+\Psi_\Lambda(\lambda_1)+\Psi_\Lambda(\lambda_2)}$ |
| Kendall's $\tau$ | $\tau = \frac{\Psi_\Lambda(\lambda_1)+\Psi_\Lambda(\lambda_2)}{\Psi_\Lambda(\lambda_1+\lambda_2)} - 1$ |
| Tail dependence | $UTD_{\hat{C}} = \frac{\Psi_\Lambda(\lambda_1)+\Psi_\Lambda(\lambda_2)-\Psi_\Lambda(\lambda_1+\lambda_2)}{\Psi_\Lambda(\max\{\lambda_1,\lambda_2\})}, \quad LTD_{\hat{C}} = 0$ |

Next, we reconsider the survival functions $u_k := \exp(-\Psi_\Lambda(\lambda_k)t_k)$ and their inverses $t_k = -\log(u_k^{1/\Psi_\Lambda(\lambda_k)})$, $k = 1,\ldots,d$. This establishes that any decreasing ordering of $t_{\pi(d)} \leq t_{\pi(d-1)} \leq \cdots \leq t_{\pi(1)}$ is one-to-one to an increasing ordering of the $u_{\pi(k)}^{1/\Psi_\Lambda(\lambda_{\pi(k)})}$ in Eq. (2). This shows that plugging in the marginal survival functions into the claimed copula yields the joint survival function and, thus, establishes the proof. □

To gain insight on the statistical properties of the derived LFC, Table 1 gives an overview of copula properties and dependence measures. Moreover, $(X_1,\ldots,X_d)$ belongs to the class of Marshall–Olkin (MO) distributions, see [7], since it fulfills the multivariate lack-of-memory property. Hence, its survival copula is known to be of extreme-value kind. A random vector $(X_1,\ldots,X_d)$ with support on $[0,\infty)^d$ on a probability space $(\Omega,\mathcal{F},\mathbb{P})$ satisfies the multivariate lack-of-memory property if

$$\mathbb{P}\left(X_{n_1} > t_1 + s,\ldots,X_{n_k} > t_k + s \mid X_{n_1} > s,\ldots,X_{n_k} > s\right) = \\ \mathbb{P}\left(X_{n_1} > t_1,\ldots,X_{n_k} > t_k\right)$$

for all $1 \leq n_1 < n_2 < \cdots < n_k \leq d$ and $s, t_1,\ldots,t_k \geq 0$, see [4, 7]. For a random vector $(X_1,\ldots,X_d)$ constructed via the inhomogeneous Lévy-frailty model, the lack-of-memory property is intuitively clear by the lack-of-memory property of the univariate trigger variables and the stationary and independent increments of the Lévy subordinator. Formally, it can be shown as follows:

$$\begin{aligned}
&\mathbb{P}\left(X_{n_1} > t_1 + s, \ldots, X_{n_k} > t_k + s \mid X_{n_1} > s, \ldots, X_{n_k} > s\right) \\
&= \frac{\mathbb{P}\left(\Lambda_{t_1+s} < E_{n_1}, \ldots, \Lambda_{t_k+s} < E_{n_k}\right)}{\mathbb{P}\left(\Lambda_s < \min\{E_{n_1}, \ldots, E_{n_k}\}\right)} \\
&= \hat{C}\left(e^{-(t_{n_1}+s)\Psi_\Lambda(\lambda_{n_1})}, \ldots, e^{-(t_{n_k}+s)\Psi_\Lambda(\lambda_{n_k})}, \underbrace{1, \ldots, 1}_{d-k \text{ times}}\right) \cdot A \\
&= \hat{C}\left(e^{-t_{n_1}\Psi_\Lambda(\lambda_{n_1})}, \ldots, e^{-t_{n_k}\Psi_\Lambda(\lambda_{n_k})}\right) \cdot \exp\left(-s\Psi_\Lambda\big(\sum_{j=1}^{k} \lambda_{\pi(j)}\big)\right) \cdot A \\
&= \hat{C}\left(e^{-t_{n_1}\Psi_\Lambda(\lambda_{n_1})}, \ldots, e^{-t_{n_k}\Psi_\Lambda(\lambda_{n_k})}\right) \\
&= \mathbb{P}\left(X_{n_1} > t_1, \ldots, X_{n_k} > t_k\right), \quad A := \left(e^{-s\Psi_\Lambda(\lambda_{n_1}+\cdots+\lambda_{n_k})}\right)^{-1}.
\end{aligned}$$

A versatile approach for constructing a random sample of $(X_1, \ldots, X_d)$ and its LFC is based on a path-wise simulation of the underlying Lévy subordinator, as given by Algorithm 1. Since most Lévy subordinators cannot be simulated continuously, a time discretization with mesh $\Delta t$ is used for simulating the increments.

## 3 Parameter Estimation

This section presents a new method for estimating the parameter vector $\theta$ of the underlying Bernstein function $\Psi_\Lambda$ and the rate parameters $\lambda_1, \ldots, \lambda_d$ of the LFC. So far, the only methodology available to estimate a high-dimensional LFC is given in [2], this approach, however, is restricted to the exchangeable case $\lambda_k \equiv 1$. For the margins $X_k \sim \mathcal{E}\left(\Psi_\Lambda\left(\lambda_k\right)\right)$ an unbiased and consistent estimator for $1/\Psi_\Lambda\left(\lambda_k\right)$ is directly given via the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} \overset{\mathbb{P}-\text{a.s.}}{\longrightarrow} \frac{1}{\Psi_\Lambda\left(\lambda_k\right)}, \qquad \text{for } n \to \infty. \tag{3}$$

Given these estimates, pseudo-samples of the LFC can be derived.

**Lemma 1** *Based on n i.i.d. samples of the LFC, an unbiased and strongly consistent estimator for* $\left(\Psi_\Lambda\left(\lambda_{s_1} + \cdots + \lambda_{s_k}\right) + 1\right)^{-1}$ *is given by*

$$1 - \frac{1}{n}\sum_{i=1}^{n} \tilde{U}^{(i)}, \tag{4}$$

*where* $\tilde{U} := \max\left\{\left(U_{s_1}\right)^{\frac{1}{\Psi_\Lambda\left(\lambda_{s_1}\right)}}, \ldots, \left(U_{s_k}\right)^{\frac{1}{\Psi_\Lambda\left(\lambda_{s_k}\right)}}\right\}$ *and* $\emptyset \neq \left\{\lambda_{s_1}, \ldots, \lambda_{s_k}\right\} \subseteq \{\lambda_1, \ldots, \lambda_d\}$.

```
Function sample_LFC (function: LevyIncSim, function: BernsteinFct, vector: $\lambda$)
    $t \leftarrow 0$
    $\Lambda \leftarrow 0$
    $\Delta t \leftarrow 0.001$
    for $k \leftarrow 1$ to $d$ do
        $E_k \leftarrow$ sample_EXP($\lambda_k$)
    end
    while true do
        $\Lambda \leftarrow \Lambda +$ LevyIncSim($\Delta t$)
        $t \leftarrow t + \Delta t$
        for $k \leftarrow 1$ to $d$ do
            if $\Lambda > E_k$ and $X_k$ not set then
                $X_k \leftarrow t$
            end
        end
        if all $X_k$ set then
            break
        end
    end
    for $k \leftarrow 1$ to $d$ do
        $U_k \leftarrow$ (exp( $-$ BernsteinFct($\lambda_k$) $\cdot X_k$ ))
    end
    return $(U_1, \dots, U_d)$
```

**Algorithm 1:** Simulation of a random sample $(U_1, \dots, U_d)$ of the LFC. The last loop can be skipped to return a sample of $(X_1, \dots, X_d)$ instead.

*Proof* From Glivenko–Cantelli, see [3, p. 20], it follows that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{\min\left\{E_{s_1}^{(i)},\dots,E_{s_k}^{(i)}\right\}\geq \Lambda_{-\log(u)}\right\}} \overset{n\to\infty}{\longrightarrow} e^{-(\lambda_{s_1}+\dots+\lambda_{s_k})\Lambda_{-\log(u)}}. \tag{5}$$

Taking the integral over $u$, applying the expectation on both sides of Eq. (5), and using conditional independence yields the estimator. Unbiasedness can easily be shown by

$$\begin{aligned}\mathbb{E}\left[1 - \frac{1}{n}\sum_{i=1}^{n} \tilde{U}^{(i)}\right] &= 1 - \int_0^1 u \cdot \Psi_\Lambda\left(\lambda_{s_1} + \dots + \lambda_{s_k}\right) u^{\Psi_\Lambda(\lambda_{s_1}+\dots+\lambda_{s_k})-1} \mathrm{d}u \\ &= \frac{1}{\Psi_\Lambda\left(\lambda_{s_1} + \dots + \lambda_{s_k}\right) + 1}.\end{aligned}$$

Strong consistency follows from the strong law of large numbers. □

Based on the estimates of $(\Psi_\Lambda\left(\lambda_{s_1} + \dots + \lambda_{s_k}\right) + 1)^{-1}$ for all non-empty $\left\{\lambda_{s_1}, \dots, \lambda_{s_k}\right\} \subseteq \{\lambda_1, \dots, \lambda_d\}$, the estimation strategy for the parameter(s) of the

Lévy subordinator (denoted $\theta$) and $\lambda_1, \ldots, \lambda_d$ is to minimize the Euclidean distance between these points of estimation and the parameterized Bernstein function; i.e.

$$\left(\hat{\theta}, \hat{\lambda}_1, \ldots, \hat{\lambda}_d\right)_n := \underset{(\theta,\lambda_1,\ldots,\lambda_d)\in\Theta}{\operatorname{argmin}} \sum_{k=1}^{2^d-1} \left(\hat{\Psi}_{S_k} - \Psi_\theta\left(\sum_{s_j \in S_k} \lambda_{s_j}\right)\right)^2, \qquad (6)$$

where $\hat{\Psi}_{S_k}$ denotes the estimated points $\left\{\Psi_A\left(\sum_{s_j\in S_k} \lambda_{s_j}\right)\right\}_{S_k\subseteq\{\lambda_1,\ldots,\lambda_d\}}$ of the Bernstein function. This is illustrated in Fig. 1.

**Numerical results**: The parameter estimation was tested via Monte Carlo simulation for 3-dim., 5-dim., and 10-dim. LFCs constructed from Poisson processes, compound Poisson subordinators, and Gamma processes. All results show a similar performance of the estimation strategy. Table 2 summarizes exemplary the mean

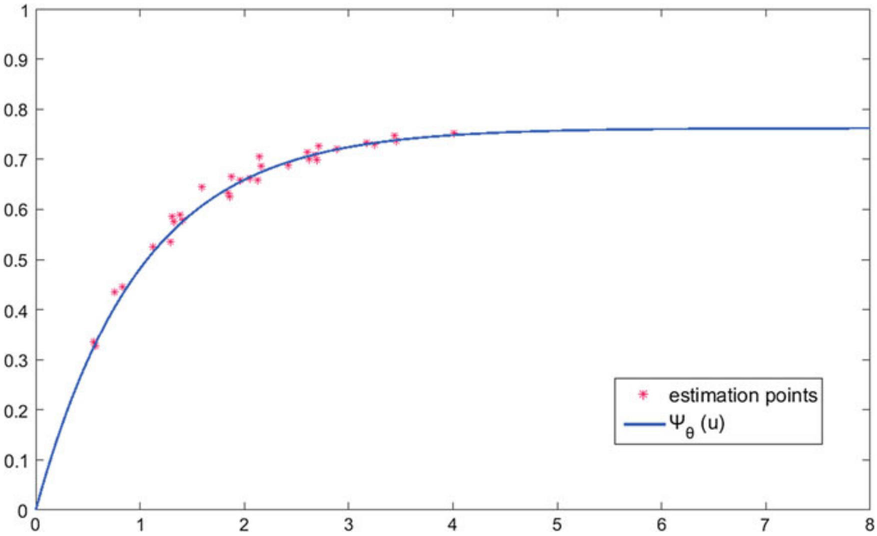


**Fig. 1** Illustration of the estimation strategy. The estimation points are marked with $*$, the Bernstein function is interpolating these. The example corresponds to the numerical values from Table 2

**Table 2** Numerical results: parameter estimation for a 5-dim. LFC wrt. a Poisson process

| | | $\theta$ | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ |
|---|---|---|---|---|---|---|---|
| True value | | 0.7622 | 0.8308 | 0.5497 | 1.2858 | 0.7537 | 0.5678 |
| $n = 50$ | Mean | 0.7804 | 0.8454 | 0.5502 | 1.3122 | 0.7670 | 0.5735 |
| | SD | 0.1468 | 0.2108 | 0.1362 | 0.3379 | 0.1930 | 0.1405 |
| $n = 100$ | Mean | 0.7722 | 0.8372 | 0.5544 | 1.2980 | 0.7612 | 0.5666 |
| | SD | 0.0978 | 0.1408 | 0.1005 | 0.2286 | 0.1276 | 0.1002 |
| $n = 1{,}000$ | Mean | 0.7612 | 0.8346 | 0.5527 | 1.2930 | 0.7565 | 0.5708 |
| | SD | 0.0301 | 0.0447 | 0.0308 | 0.0671 | 0.0403 | 0.0299 |

and standard deviation of 1,000 iterations for the 5-dim. LFC built from a Poisson process. In this particular example it seems that the method slightly overestimate the parameter values. This, however, is not confirmed by other examples.

## 4 Conclusion

A new family of non-exchangeable extreme-value copulas was derived and analyzed in some detail. The objective was to construct a copula that is intuitive, flexible, and tractable while being non-exchangeable and satisfying the extreme-value property. Such dependence models are of relevance for at least two reasons. First, most well-known copulas are exchangeable, while real-world applications often impose the need for non-exchangeable structures. Second, joint extreme events can have a massive impact in finance, insurance, or environmental science. The developed family of LFC can provide an appropriate model for the analysis of such situations. Furthermore, a new estimation strategy for the parameters of the model was established.

## References

1. Bertoin J (1999) Subordinators: Examples and Applications. Springer, Berlin, Heidelberg
2. Hering C, Mai J-F (2012) Moment-based estimation of extendible Marshall-Olkin copulas. Metrika 75:601–620
3. Loève M (1963) Probability theory. The university series in higher mathematics. Van Nostrand, Princeton NJ
4. Mai J-F (2014) Multivariate exponential distributions with latent factor structure and related topics. Habilitation thesis, Technische Universität München, München
5. Mai J-F, Scherer M (2009) Lévy-frailty copulas. J Multivar Anal 100:1567–1585
6. Mai J-F, Scherer M (2014) Financial engineering with copulas explained. Palgrave Macmillan, Basingstok
7. Marshall AW, Olkin I (1967) A multivariate exponential distribution. J Am Stat Assoc 31:30–44

# A Perceptron Classifier and Corresponding Probabilities

**Bernd-Jürgen Falkowski**

**Abstract** In this paper a fault tolerant probabilistic kernel version with smoothing parameter of Minsky's perceptron classifier for more than two classes is sketched. Moreover a probabilistic interpretation of the output is exhibited. The price one has to pay for this improvement appears in the non-determinism of the algorithm. Nevertheless an efficient implementation using for example Java concurrent programming and suitable hardware is shown to be possible. Encouraging preliminary experimental results are presented.

**Keywords** Perceptron · Classifier for more than 2 classes · Bayes decision

## 1 Introduction

Recently the analysis of Big Data has become increasingly important. Indeed, implementations of classifying and in particular ranking algorithms have been effected in order to perform such diverse tasks as assessing the creditworthiness of banking customers, supporting medical doctors in their diagnoses of patients, ranking drivers according to their driving behaviour (as made possible by modern navigation systems) or establishing recommender systems for online shops. Here an improvement of an old classification algorithm is sketched out that is appealing from an aesthetic point of view due to its elegant simplicity, whilst in addition preliminary experimental results indicate that it might well also be suitable for commercial applications.

In Sect. 2 the original algorithm is described together with a generalization for more than 2 classes and a geometric interpretation is given. In Sect. 3.1 the kernel trick is exhibited whilst in Sect. 3.2 the new algorithm is sketched out. In Sect. 4 a probabilistic interpretation of the decision procedure is given. Preliminary experimental results in Sect. 5 and a conclusion and outlook in Sect. 6 end the paper.

B.-J. Falkowski (✉)
Fachhochschule für Ökonomie und Management FOM, Karlstrasse 2,
86150 Augsburg, Germany
e-mail: bernd.falkowski@fh-stralsund.de

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_27

## 2 The Perceptron

In their seminal work [9] Minsky and Papert describe a perceptron as a simple classifier by means of a linear threshold function as follows.

**Definition 1** Let $\Phi := \{\phi_1, \phi_2, \ldots, \phi_m\}$ be a family of (generalized) predicates (in general real valued functions defined on some set of objects). Then the truth-valued function $\psi$ (predicate) is a linear threshold function with respect to $\Phi$ if there exists a real number $\theta$ and coefficients $\alpha(\phi_1)$, $\alpha(\phi_2)$, ..., $\alpha(\phi_m)$ such that $\psi(x)$ = true if and only if $\sum_{i=1}^{m} \alpha(\phi_i)\phi_i(x) > \theta$. Any predicate that can be defined in this way is said to belong to $L(\Phi)$.

Now suppose that two disjoint sets of objects $S^+$ and $S^-$ and a family of generalized predicates $\Phi$ on $S = S^+ \cup S^-$ are given.Then one would like to construct a predicate $\psi$ in $L(\Phi)$ such that $\psi(x)$ = true if and only if $x \in S^+$, in other words one would like to construct a $\psi$ in $L(\Phi)$ that separates $S^+$ and $S^-$.

As shown by Minsky and Papert this can be done using the following simple program (the Perceptron Learning Algorithm, **PLA**), in which the convenient scalar product notation $\mathbf{A} \cdot \mathbf{\Phi}(x)$ instead of $\sum_{i=1}^{m} \alpha(\phi_i)\phi_i(x)$ is used and $\mathbf{A} := (\alpha(\phi_1), \ldots, \alpha(\phi_{\mathrm{m}}))$ and $\mathbf{\Phi}(x) := (\phi_1(x), \ldots, \phi_m(x))$ are considered as elements of $\mathbb{R}^{\mathrm{m}}$, if a solution exists. (It is instructive to note here that the basic geometric concepts of length and angle may be described in purely algebraic terms using the scalar product. Taking this into account and generalizing to higher dimensions the solution may thus be considered in geometrical terms as a separating hyperplane. However, the set $S$ is not required to carry a vector space structure although in practical applications this will often be the case.).

```
Start      Choose any value for $\mathbf{A}$, $\theta$.
Test       If $x \in S^+$ and $\mathbf{A} \cdot \mathbf{\Phi}(x) > \theta$ go to Test.
           If $x \in S^+$ and $\mathbf{A} \cdot \mathbf{\Phi}(x) \leq \theta$ go to Add.
           If $x \in S^-$ and $\mathbf{A} \cdot \mathbf{\Phi}(x) < \theta$ go to Test.
           If $x \in S^-$ and $\mathbf{A} \cdot \mathbf{\Phi}(x) \geq \theta$ go to Subtract .
Add        Replace $\mathbf{A}$ by $\mathbf{A} + \mathbf{\Phi}(x)$ and $\theta$ by $\theta - 1$. Go to Test.
Subtract   Replace $\mathbf{A}$ by $\mathbf{A} - \mathbf{\Phi}(x)$ and $\theta$ by $\theta + 1$. Go to Test.
```

Having found a suitable vector $\mathbf{A}^\star$ and a scalar $\theta^\star$ the decision procedure for classification is given by:
Decide $x \in S^+$ if and only if $\mathbf{A}^\star \cdot \mathbf{\Phi}(x) > \theta^\star$.
If there exists a more general partition of $S = \bigcup_{i=1}^{q} S_i$, then one can still construct a suitable classifier as follows. Given $\mathbf{\Phi}$ as above, find a vector $\mathbf{A}^\star := (\mathbf{A}_1^\star, \mathbf{A}_2^\star, \ldots, \mathbf{A}_{\mathrm{q}}^\star)$ and a number $\theta^*$ such that $\mathbf{A}_{\mathrm{i}}^\star \cdot \mathbf{\Phi}(x) > \mathbf{A}_{\mathrm{j}}^\star \cdot \mathbf{\Phi}(x) + \theta^*$ for all $j \neq i$ if and only if $x \in S_i$. This problem can be reduced to the one described above by the following definition.

**Definition 2** Define a new vector $\mathbf{\Phi}_{\mathrm{ij}} := (0, \ldots, 0, \mathbf{\Phi}(x), 0, \ldots, 0, -\mathbf{\Phi}(x), 0..., 0)$ containing $\mathbf{\Phi}(x)$ in the i-th place and $\mathbf{\Phi}(x)$ in the j-th place.

Indeed, this definition leads to the following program.

```
Start      Choose any value for $\mathbf{A}$, $\theta$.
Test       If $x \in S_i$ and $\mathbf{A} \cdot \mathbf{\Phi}_{\mathrm{ij}}(x) > \theta$ go to Test.
           If $x \in S_i$ and $\mathbf{A} \cdot \mathbf{\Phi}_{\mathrm{ij}}(x) \leq \theta$ go to Add.
           If $x \in S_j$ and $\mathbf{A} \cdot \mathbf{\Phi}_{\mathrm{ij}}(x) < \theta$ go to Test.
           If $x \in S_j$ and $\mathbf{A} \cdot \mathbf{\Phi}_{\mathrm{ij}}(x) \geq \theta$ go to Subtract.
Add        Replace $\mathbf{A}$ by $\mathbf{A} + \mathbf{\Phi}_{\mathrm{ij}}(x)$ and $\theta$ by $\theta - 1$. Go to Test.
Subtract   Replace $\mathbf{A}$ by $\mathbf{A} - \mathbf{\Phi}_{\mathrm{ij}}(x)$ and $\theta$ by $\theta + 1$. Go to Test.
```

Note, in order to avoid confusion, that, by abuse of notation, the same $\mathbf{A}$, $\theta$ as above have been used.
Note also that having found a suitable $\mathbf{A}^{\star}$ and a scalar $\theta^{\star}$ the decision procedure for classification is of course:
Decide $x \in S_i$ if and only if $\mathbf{A}^{\star} \cdot \mathbf{\Phi}_{\mathrm{ij}}(x) > \theta^{\star}$ for all $j \neq i$.

The interesting point about this program is the fact that, as already noted by Minsky and Papert, a straightforward error-correcting feedback results in a correct algorithm. Of course, the required existence of a solution is by no means guaranteed in general although, if suitable predicates are used, in many cases a solution can be found, cf. [3]. However, nowadays good generalization properties of the perceptron are of paramount importance and hence it is often preferable to admit a solution that does not separate the $S_i$ completely, see also [2].

## 3 Kernel Learning

In order to avoid having to deal explicitly with extremely high dimensional spaces that lead to unacceptable CPU times one applies the so-called kernel trick, cf. [11].

### *3.1 Positive Definite Kernels*

If above one starts with the zero vector $\mathbf{0}$ for $\mathbf{A}$ in order to avoid technical complications then it is easily seen that finally $\mathbf{A}$ will have the form

$$\mathbf{A} = (\sum_{\mathrm{k=1}}^{\mathrm{m}} b_{1k} \mathbf{\Phi}(x_k), \sum_{k=1)}^{m} b_{2k} \mathbf{\Phi}(x_k), \ldots, \sum_{k=1}^{m} b_{mk} \mathbf{\Phi}(x_k))$$

for some coefficients $b_{jk}$. Hence $\mathbf{A}$ may equally well be described by the vector

$$\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{\mathrm{m}})$$

in terms of the $b_{ij}$. Moreover

$$\mathbf{A} \cdot \mathbf{\Phi}_{\mathrm{ij}}(x) = \sum_{k=1}^{m} (b_{ik} - b_{jk}) \mathbf{\Phi}(x_k) \cdot \mathbf{\Phi}(x) = \sum_{k=1}^{m} (b_{ik} - b_{jk}) K(x_k, x)$$

say, and the update operation is given by

$$\mathbf{A} + \mathbf{\Phi}_{\mathrm{ij}}(x_s) = (\sum_{k=1}^{m} c_{1k} \mathbf{\Phi}(x_k), \sum_{k=1}^{m} c_{2k} \mathbf{\Phi}(x_k), \dots, \sum_{k=1}^{m} c_{mk} \mathbf{\Phi}(x_k))$$

where $x_s \in S_i$ is assumed, and $c_{is} = b_{is} + 1$, $c_{js} = b_{js} - 1$ and elsewhere $c_{rl} = b_{rl}$. This update operation may of course be described entirely in terms of the vectors $\mathbf{b}$ and $\mathbf{d_{ij}}$= $(\mathbf{0}, \dots, \mathbf{0}, \mathbf{d_i}, \mathbf{0}, \dots, \mathbf{0}, \mathbf{d_j}, \mathbf{0}, \dots, \mathbf{0})$ that has a 1 as the $d_{is}$ entry, a minus 1 as the $d_{js}$ entry and zeroes elsewhere by

$$\mathbf{b} := \mathbf{b} + \mathbf{d}_{\mathrm{ij}}$$

Thus one is lead to the following definition that formalizes the foregoing considerations.

**Definition 3** A real-valued function $K : S \times S \to \mathbb{R}$ is called a positive definite kernel if for all choices of n, and $x_1, x_2, , x_n \in S$ the matrix with entries $K(x_i, x_j)$ is symmetric and positive definite.

Given such a kernel an embedding $\mathbf{\Phi}$ of $S$ in a vector space $H = \mathbb{R}^S$ (the space of functions from $S$ to $\mathbb{R}$) may always be constructed by setting $\mathbf{\Phi}(x) := \mathrm{K}(., x)$, considering functions $f = \sum_{j=1}^{m} \gamma_j K(., x_j)$, and defining addition of such functions and multiplication of such a function by a scalar pointwise. If the inner product is defined by $< \mathbf{\Phi}(x), \mathbf{\Phi}(y) >_H := K(x, y)$ and extended by linearity, then a Hilbert space H (the Reproducing Kernel Hilbert space) is obtained by completion as usual, see e.g. [11]. Hence a positive definite kernel is seen to be the abstract version of a scalar product. Of course, given a positive definite kernel one may now discard the set of predicates entirely and arrive at the following algorithm constructing a separating hyperplane, which is obtained from Minsky's original version.

**Start** Choose any value for $\theta$. and set $b_{rl} = 0$ for all r,l.

**Test** If $x_s \in S_i$ and $\sum_{k=1}^{m} (b_{ik} - b_{jk}) K(x_k, x_s) > \theta$ go to **Test**.
If $x_s \in S_i$ and $\sum_{k=1}^{m} (b_{ik} - b_{jk}) K(x_k, x_s) \leq \theta$ go to **Add**.
If $x_s \in S_j$ and $\sum_{k=1}^{m} (b_{ik} - b_{jk}) K(x_k, x_s) < \theta$ go to **Test**.
If $x_s \in S_j$ and $\sum_{k=1}^{m} (b_{ik} - b_{jk}) K(x_k, x_s) \geq \theta$ go to **Subtract**.

**Add** Replace $\mathbf{b}$ by $\mathbf{b} + \mathbf{d}_{\mathrm{ij}}$ and $\theta$ by $\theta - 1$. Go to **Test**.

**Subtract** Replace $\mathbf{b}$ by $\mathbf{b} - \mathbf{d}_{\mathrm{ij}}$ and $\theta$ by $\theta + 1$. Go to **Test**.

The above program again computes a weight vector

$$\mathbf{b}^\star = (\mathbf{b}_1^\star, \mathbf{b}_2^\star, \dots, \mathbf{b}_\mathrm{m}^\star)$$

and a scalar $\theta^\star$ such that the decision procedure is given by:
Decide $x_s \in S_i$ if and only if $\sum_{k=1}^m (b_{ik}^\star - b_{jk}^\star) K(x_k, x_s) > \theta^\star$ for all $j \neq i$.

### 3.2 *The Optimal Separating Hyperplane*

The algorithm in Sect. 3.1 computes a separating hyperplane in the reproducing kernel Hilbert space, if it exists. However, in this case it is desirable to arrive at a hyperplane that is optimal in the sense that the minimum distance of any point from the plane is maximal, cf. [12]. This can also be achieved by simple feedback if one tests the "worst classified element" instead of an arbitrary one. Details are given in the description of the Krauth/Mezard algorithm, cf. [8]. In fact this amounts to minimizing the (square of the) norm of the weight vector

$$\mathbf{A} = (\sum_{\mathrm{k=1}}^{\mathrm{m}} b_{1k} \boldsymbol{\Phi}(x_k), \sum_{k=1)}^{m} b_{2k} \boldsymbol{\Phi}(x_k), \dots, \sum_{k=1}^{m} b_{mk} \boldsymbol{\Phi}(x_k))$$

given by

$$\|\mathbf{A}\|^2 = \sum_{\mathrm{k=1}}^{\mathrm{q}} \sum_{\mathrm{i=1}}^{\mathrm{m}} \sum_{\mathrm{j=1}}^{\mathrm{m}} b_{ki} b_{kj} \boldsymbol{\Phi}(x_i) \cdot \boldsymbol{\Phi}(x_j) = \sum_{k=1}^{q} \sum_{i=1}^{m} \sum_{j=1}^{m} b_{ki} b_{kj} K(x_i, x_j)$$

where q is the number of classes, as can easily be seen. A detailed proof may be found in [12].

This motivates the introduction of the target function $E(D) + \lambda \|\mathbf{A}\|^2$, where $\lambda$ is a smoothing parameter to be determined experimentally.

Here $E(D)$ denotes the empirical risk which may be the number (or more generally cost) of errors whilst the second term describes the structural risk. Of course, now the number of errors will not be zero in general at the minimum of the target function. Nevertheless the **PLA** can be modified so as to be still applicable by endowing it with a ratchet. The resulting algorithm, based on a kernel version of the Pocket Algorithm, cf. [6, 7], is a probabilistic one. A correctness proof based on the representer theorem, cf. [4, 10] can be constructed.

## 4 Probabilistic Interpretation of the Decision Procedure

The decision procedure described above may under certain conditions be interpreted as Bayes decision. Indeed, assume that the class conditional densities belong to the family of exponential distributions (which includes a number of well-known distributions) of the general form

$$p(\mathbf{\Phi}(\mathrm{x})|\mathbf{x} \in \mathrm{S_i}) = \exp(\mathrm{B}(\mathbf{e}_\mathrm{i}) + \mathrm{C}(\mathbf{\Phi}(\mathrm{x})) + < \mathbf{e}_\mathrm{i}, \mathbf{\Phi}(\mathrm{x}) >)$$

where $\mathbf{x}$ is now a vector in some Euclidean space and the $\mathbf{e}_\mathrm{i}$ are parameter vectors. Then the posterior probabilities can be computed using Bayes theorem.

**Theorem 1** *If the class conditional densities belong to the family of exponential distributions and* $B(\mathbf{e}_\mathrm{i})+\ln(\mathrm{P}(\mathrm{S_i}))$ *is independent of i the decision procedures derived above will give the Bayes decision.*

*Proof* The posterior probabilities can be computed using Bayes theorem as

$$p(\mathbf{x} \in \mathrm{S_i}|\mathbf{\Phi}(\mathrm{x}) =$$

$$\frac{p(\mathbf{\Phi}(\mathrm{x})|\mathbf{x} \in \mathrm{S_i}) * \mathrm{P}(\mathrm{S_i})}{\sum_j p(\mathbf{\Phi}(\mathrm{x})|\mathbf{x} \in \mathrm{S_j}) * \mathrm{P}(\mathrm{S_j})} =$$

$$\frac{\exp(B(\mathbf{e}_\mathrm{i}) + \mathrm{C}(\mathbf{\Phi}(\mathrm{x})) + < \mathbf{e}_\mathrm{i}, \mathbf{\Phi}(\mathrm{x}) >) * \mathrm{P}(\mathrm{S_i})}{\sum_j \exp(B(\mathbf{e}_j) + \mathrm{C}(\mathbf{\Phi}(\mathrm{x})) + < \mathbf{e}_\mathrm{j}, \mathbf{\Phi}(\mathrm{x}) >) * \mathrm{P}(\mathrm{S_j})} =$$

$$\frac{\exp(B(\mathbf{e}_\mathrm{i}) + < \mathbf{e}_\mathrm{i}, \mathbf{\Phi}(\mathrm{x}) > + \ln(\mathrm{P}(\mathrm{S_i}))}{\sum_j \exp(B(\mathbf{e}_\mathrm{j}) + < \mathbf{e}_\mathrm{j}, \mathbf{\Phi}(\mathrm{x}) > + \ln(\mathrm{P}(\mathrm{S_j}))} =$$

Setting $\mathbf{A}_\mathrm{i} := \mathbf{e}_\mathrm{i}$ and $-\theta := B(\mathbf{e}_\mathrm{i})+\ln(\mathrm{P}(\mathrm{S_i}))$ one obtains

$$p(\mathbf{x} \in \mathrm{S_i}|\mathbf{\Phi}(\mathrm{x})) = \frac{\exp(\mathrm{a_i})}{\sum_\mathrm{j} \exp(\mathrm{a_j})}$$

with

$$a_i =< \mathbf{A}_\mathrm{i}, \mathbf{\Phi}(\mathrm{x}) > -\theta$$

Hence it becomes clear that, provided that the assumed class conditional density is appropriate and that the above substitutions are justified, deciding that an $\mathbf{x}$ belongs to $S_i$ if $a_i$ is maximal is also the Bayes decision since the a posteriori probability is maximal in this case.
The kernel version of $a_i$, say $k_i$ is then given by

$$k_i = \sum_{k=1}^{m} b_{ik} K(\mathbf{x}_\mathrm{k}, \mathbf{x}) - \theta$$

and again deciding that an $\mathbf{x}$ belongs to $S_i$ if $b_i$ is maximal is the Bayes decision. □

In this context also note that the function given in the equation above describes a generalization of the logistic sigmoid activation function which is known as the normalized exponential or softmax activation function: This function represents a smooth version of the winner-takes-all activation model, For further details see e.g. [1].

## 5 Preliminary Experimental Results

In order to get some information on the practical value of the fault tolerant kernel version of the pocket algorithm using a smoothing parameter mentioned at the end of Sect. 3.2 some performance and functionality tests were conducted using some real life data. The experiments were carried out with 2 * 5964 sample vectors (constituting a training and a validation set) provided in anonymous form by a German financial institution. The customers had been divided into 2 preference classes and in both sets there were exactly 123 bad customers. The experiments were conducted on a commercially available PC (QuadCore, $4 \times 2$, 4 GHz processors, 8 GB RAM). The operating system was Windows 7 and the current Java version was used applying concurrent programming, cf. [5]. Using a cubic kernel and 20 000 iterations of the main loop took about 2 h of CPU time. The generalization properties (as judged using the validation set) compared favourably with standard methods like logistic regression.

Of course, it must be admitted that there were 2 classes only available for these tests. Moreover whilst training results with an RBF (Radial Basis Function) kernel were very good, generalization properties turned out to be rather poor for this kernel (probably due to overfitting in view of the increased Vapnik-Chervonenkis bound, see [13]). As a consequence no reliable information concerning the practical use of the algorithm is available although the results described above seem encouraging.

## 6 Conclusion and Outlook

An elegant and compact probabilistic algorithm relying on straightforward error correcting feedback has been sketched (derived essentially from Minsky's original perceptron learning algorithm) and a correctness proof has been hinted at. A probabilistic interpretation of the output has been provided using Bayes' theorem. Good generalization properties due to the introduction of a smoothing parameter appear to be likely as indicated by preliminary experimental results exhibited in Sect. 5. Whilst the currently needed CPU times to execute the kernel version of the algorithm are still in the region of several hours the possibility of parallelization (as demonstrated with Java concurrent programming in a test version) allows significant improvements if suitably sophisticated hardware is employed. Thus it seems that an algorithm has been obtained that is not only very appealing from an aesthetic point of view but could also quite successfully be used in commercial and academic applications. Of course, in order to prove the commercial viability extensive experimental work is still necessary.

Nevertheless, it seems rather remarkable that an old algorithm, that originally was probably mainly of academic interest, can now be implemented using sophisticated hardware and concurrent programming techniques in such a way that it retains most of its inherent simplicity whilst also being of interest for practical applications.

**Acknowledgments** The author is indebted to M. Stern for help with some problems concerning the Java system.

## References

1. Bishop CM (2006) Pattern recognition and machine learning. Springer
2. Block HD, Levin SA (1970) On the boundedness of an iterative procedure for solving a system of linear inequalities. In: Proceedings of the AMS
3. Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans Electron Comput 14
4. Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmanns D (eds) Advances in large margin classifiers. MIT Press
5. Falkowski B-J (2008) Parallel implementation of certain neural network algorithms. In: Ruan D, Montero J, Lu J, Martinez L, D hondt P, Kerre E (eds) Computational intelligence in decision and control. Proceedings of the 8th international FLINS conference. World Scientific
6. Falkowski B-J (2015) Minsky revisited, fault tolerant perceptron learning with good generalization properties. In: Miller L (ed) Proceedings of 30th international conference on computers and their applications (CATA 2015). ISCA Publications
7. Gallant SI (1990) Perceptron-based learning algorithms. IEEE Trans Neural Netw I(2) (1990)
8. Krauth W, Mezard M (1987) Learning algorithms with optimal stability in neural networks. J Phys A: Math Gen 20
9. Minsky ML, Papert S (1990) Perceptrons, Expanded edn. MIT Press
10. Schölkopf B, Herbrich R, Smola AJ (2001) A generalized representer theorem. Computational learning theory. LNCS, vol 2111
11. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press
12. Smola AJ, Bartlett PL, Schölkopf B, Schuurmanns D (2000) Introduction to large margin classifiers. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmanns D (eds) Advances in large margin classifiers. MIT Press
13. Vapnik VN (1998) Statistical learning theory. Wiley

# Fuzzy Signals Fed to Gaussian Channels

**Laura Franzoi and Andrea Sgarro**

**Abstract** We add fuzziness to the signals sent through a continuous Gaussian transmission channel: fuzzy signals are modeled by means of triangular fuzzy numbers. Our approach is mixed, fuzzy/stochastic: below we do not call into question the probabilistic nature of the channel, and fuzziness will concern only the channel inputs. We argue that fuzziness is an adequate choice when one cannot control crisply each signal fed to the channel. Using the fuzzy arithmetic of interactive fuzzy numbers, we explore the impact that a fuzziness constraint has on channel capacity; in our model we are ready to tolerate a given fuzziness error $F$. We take the chance to put forward a remarkable case of "irrelevance" in fuzzy arithmetic.

## 1 Introduction

In a finite or discrete setting, a fuzzy or possibilistic approach to information and coding theories has been pursued starting with [11] up to [2], going through [1]. Below, we shall consider a standard *continuous* Gaussian channel with additive white noise; our aim is to compute the modified (diminished) channel capacity when channel inputs are allowed to be fuzzy, more precisely are triangular *fuzzy signals* $S_i$, $1 \le i \le n$, which have to verify the power constraint $\frac{1}{n}\sum_i S_i^2 \le P$ in an approximate sense to be made precise below. The crisp positive number $P$ is the given *energy power*, cf. below Sects. 3 and 4, where our model of data transmission is described and vindicated. While Sect. 2 is devoted to technical preliminaries, Sect. 3, after some computations in fuzzy arithmetic, deals with what should be an adequate information-theoretic notion of *fuzziness error*, as a sort of counterpart to the stan-

L. Franzoi (✉)
University of Bucharest, Str. Academiei 14, 010014 Bucharest, Romania
e-mail: laura.franzoi@gmail.com

A. Sgarro
Department of Mathematics and Geosciences, University of Trieste,
P.le Europa, 1, 34100 Trieste, Italy
e-mail: sgarro@units.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_28

dard probabilistic notion of *decoding error*. The final Sect. 4 is information-theoretic and relies heavily on [5]; however, we have tried to present a paper which would be as self-contained as possible. As a standard reference to fuzzy sets cf. e.g. [7].

## 2 Preliminaries and Technicalities

In the approach to fuzzy arithmetic which we are taking, an approach which was pioneered in [6] and which is largely followed today, cf. e.g. [3, 4, 8, 9, 12, 13], a fuzzy $n$-tuple $\underline{X} = X_1, \dots, X_n$ is defined by giving its *distribution function* $f(\underline{x}) : \mathbb{R}^n \to [0, 1]$, $\underline{x} \doteq x_1, \dots, x_n$, where the equation $f(\underline{x}) = 1$ admits of at least one solution, usually, as happens below, of *exactly* one solution. The advantage of the joint distribution approach is not only its generality, but first and foremost the fact that the computation rules are *the same*[1] as those in the usual arithmetic of crisp real numbers. For $n = 1$ our *fuzzy numbers* $X$ of distribution function $f(x) \doteq X(x)$ will be triangular-like, i.e. their continuous distribution function increases from 0 to 1 on the interval $[a, b]$ and then decreases from 1 to 0 on the interval $[b, c]$, $a \le b \le c$. One of the two monotone components might be lacking; *crisp numbers* $b$ can be identified with fuzzy numbers whose distribution function $X(x)$ is 1 in $b$, else is 0. In this paper the *fuzziness* of number $X$ will be identified with the *area* $\int_a^c X(x)dx$.

If $\psi : \mathbb{R}^n \to \mathbb{R}$ is a function, the distribution function $Z(z)$ of the fuzzy quantity $Z \doteq \psi(X_1, \dots, X_n)$ is given[2] by:

$$Z(z) = \max_{\underline{x}:\, \psi(\underline{x})=z} f(\underline{x}) \tag{1}$$

with $Z(z) \doteq 0$ when the maximization set is void. A relevant case is when $\psi(x_1, \dots, x_n) = x_i$: this allows one to obtain the $n$ distribution functions of the *marginal* fuzzy quantities $X_1, \dots, X_n$. Another relevant case is when $n = 2$ and so $\psi(x, y)$ is a binary operation $x \circ y$ on fuzzy quantities $X$ and $Y$ of joint distribution function $f(x, y)$, $Z = \psi(X, Y) \doteq X \circ Y$.

$$Z(z) = \max_{x,y:\, x\circ y=z} f(x, y) \tag{2}$$

Let us take $n$ fuzzy quantities $\underline{X} \doteq X_1, \dots, X_n$. An *admissible* $n$-dimensional *joint distribution* $f(x_1, \dots, x_n)$ can be derived by just taking minima $\wedge$ and setting it

[1] Actually $f(x_1, \dots, x_n) = g(x_1, \dots, x_n)$ is an identity for crisp numbers $x_1, \dots, x_n$ if and only if the two fuzzy quantities $Z_1 \doteq f(X_1, \dots, X_n)$ and $Z_2 \doteq g(X_1, \dots, X_n)$ are deterministically equal whatever the joint distribution of the n fuzzy quantities $X_1, \dots, X_n$; deterministic equality is formally defined in this Section. Cf. e.g. [9] for a discussion of this straightforward but important result, called there the *Montecatini lemma*.

[2] Our functions below are quite well-behaved, and so we write directly *maxima* rather than *suprema*. With a slight imprecision, when the support of a function (the set where the function is positive) is an interval, the interval will be closed anyway.

equal to $X_1(x_1) \wedge \ldots \wedge X_n(x_n)$. In this case, one says that the $n$ random quantities are *non-interactive*; the joint distribution is *admissible* in the sense that the $n$ distribution functions $X_i(x)$ are re-obtained as marginal distributions, $1 \le i \le n$. As well-known, cf. e.g. [10], non-interactivity is (rightly) seen as a fuzzy analogue of *probabilistic independence* for $n$ random variables. As equally well-known and anyway soon proven, for given $X_1, \ldots, X_n$, the non-interactive joint distribution is the unique *maximum* in the partially ordered set of admissible joint distributions $f(\underline{x})$, where $f_1 \le f_2$ if the inequality $f_1(\underline{x}) \le f_2(\underline{x})$ holds whatever the $n$-dimensional argument $\underline{x}$. Another relevant admissible joint distribution of $\underline{X} \doteq X_1 \ldots X_n$ and $\underline{Y} \doteq Y_1 \ldots Y_n$ is *deterministic equality*: in this case one has to assume *equidistribution* $\underline{X} \simeq \underline{Y}$, i.e. equality of the two respective distribution functions $f$ and $g$, $f(\underline{x}) = g(\underline{x})$ for all $x$, and the joint distribution is 0 unless $\underline{x} = \underline{y}$, in which case it is equal to $f(\underline{x}) = g(\underline{x}) = f(\underline{y}) = g(\underline{y})$. In our approach to fuzzy arithmetic, only under deterministic equality $\underline{X} = \underline{Y}$ one of the two symbols $\underline{X}$ or $\underline{Y}$ is disposable, while equidistribution $\underline{X} \simeq \underline{Y}$ in itself is *not* enough.

To ease reference we find it convenient to mention explicitly a few well-known and soon proven facts; below we say that a fuzzy number is a triangle $(a, b, c)$ when its distribution function increases *linearly* from 0 to 1 on $[a, b]$ and decreases *linearly* from 1 to 0 on $[b, c]$.

1. If $X_1, \ldots, X_n$ are non-interactive, so are any deterministic functions thereof, $\psi_1(X_1), \ldots, \psi_n(X_n)$.
2. If $Y = \alpha X + \beta$, $\alpha > 0$, the distribution function $Y(y) = X(\frac{y-\beta}{\alpha})$ has the same "shape" as $X(x)$, only translated and re-scaled. The *non-interactive* sum of two triangles $(a_1, b_1, c_1)$ and $(a_2, b_2, c_2)$ is $(a_1 + a_2, b_1 + b_2, c_1 + c_2, )$.
3. If $X(-x) \le X(x)$ for any $x \ge 0$, e.g. if $X(x)$ is symmetric around 0, the distribution function of the absolute value $|X|$ equals $X(x)$ for $x \ge 0$, else is 0.
4. Let $X$ and $Y$ be interactive with joint distribution function $f(x, y)$; if $Z \doteq X \circ Y$ has the same distribution function $Z(z)$ also under non-interactivity, then the distribution function of $Z$ remains the same under *any* joint distribution $f_1 > f$ in the partially ordered set of admissible joint distributions.

Point 4 is a special case of *irrelevance*, a convenient notion introduced and discussed in [8, 12, 13]; a remarkable case of irrelevance will be Theorem 1 below. The following lemma is stated explicitly, and represents a time-honored *ante litteram* case of irrelevance involving non-interactivity and deterministic equality. We give it in a rather general form, cf. [8]; we are not assuming commutativity of the operation $x \circ y$, which however has to be *order-preserving*: $x \le u,\ y \le v$ implies $x \circ y \le u \circ v$.

**Lemma 1** *Take $X$ and $Y$ non-interactive and equidistributed, with $X(x) = Y(x)$ a (possibly weakly) concave function over its connected support. Take an order-preserving operation $x \circ y$ such that $h(x) \doteq x \circ x$ is a continuous function of its argument. Then one has $X \circ Y \simeq X \circ X$.*

*Proof* The function $h(x)$ is itself non-decreasing on the support. If $u < v$, one has $h(u) \doteq u \circ u \leq u \circ v \leq v \circ v \doteq h(v)$ since $\circ$ is order-preserving, and so, by continuity of $h(x)$, there is a value $x = \alpha u + (1-\alpha)v$ between $u$ and $v$ $(0 \leq \alpha \leq 1)$ such that $h(x) \doteq u \circ v$; then by concavity $X(x) = Y(x) \geq \alpha X(u) + (1-\alpha)X(v) \geq X(u) \wedge X(v) = X(u) \wedge Y(v)$. Use also point 4 above.

## 3 Fuzziness Error

The fuzzy signal $S_i$ will be modeled by a linear triangular fuzzy number $(s_i - c, s_i, s_i + c) = (s_i - 1, s_i, s_i + 1)$, setting without real restriction[3] its area $c$, i.e. its *fuzziness*, equal to 1 independent of $i$. In (3) the triangular number $X_i = (-1, 0, +1)$ represents the "unit fuzziness" summed to each crisp signal $s_i$. The fuzzy power constraint is soon found to be:

$$\sum_i S_i^2 \doteq Z + \sum_i s_i^2 = \sum_i X_i^2 + 2\sum_i |s_i| \cdot X_i + \sum_i s_i^2 \ \leq \ nP \qquad (3)$$

We wrote absolute values because if $s_i$ is negative the triangle $s_i X_i$ is $(s_i, 0, -s_i)$. In (3) the fuzzy term $Z$ with $Z(0) = 1$ represents the "total fuzziness" summed to the crisp sum of squares $\sum_i s_i^2$.
Let us deal with the distribution function $Z(z)$ of $Z = \sum_i X_i^2 + 2\sum_i |s_i| \cdot X_i$, cf. (3). Actually, we will be interested only in non-negative values $z$ and in the fuzzy quantity $Z^+ \doteq |Z|$ rather than $Z$, whose distribution function, using point 3, Sect. 2, turns out to be $Z^+(z) = Z(z)$ for $z \geq 0$, else $Z^+(z) \doteq 0$. So, referring again to point 3:

$$Z^+ = \sum_i X_i^2 + 2\sum_i |s_i| X_i^+ \,, \quad X_i^+ \doteq |X_i|$$

In practice, with respect to (3) the original triangles $(-1, 0, +1)$ are replaced by "degenerate" triangles $(0, 0, +1)$. The fuzzy term $2\sum_i |s_i| X_i^+$, cf. point 2 above, is the triangle $(0, 0, 2ns^{(n)})$ where we set $s^{(n)} \doteq \frac{1}{n}\sum_i |s_i|$.

**Theorem 1** *The distribution function of the fuzzy number* $Z^+$ *is*

$$Z^+(z) = 1 + s^{(n)} - \sqrt{\frac{z}{n} + [s^{(n)}]^2} \,, \quad z \in \big[0, n\big(1 + 2s^{(n)}\big)\big]$$

*Proof* Fix $i$; $X_i^2 + 2|s_i|X_i^+$ is a deterministic function of $X_i$, and its distribution function is soon found to be $1 + s_i - \sqrt{z + s_i^2}$ , $z \in \big[0, 1 + 2s_i\big]$, cf. (1). As for the whole sum of $n$ terms, one can use an induction on $n$, resorting to the Addendum at

[3] Else, one might re-scale each $S_i$ of fuzziness $c$, $c > 0$, to $S_i/c$ of fuzziness 1, and consequently re-scale $P$ and $N$ (cf. Sect. 4) dividing them by $c^2$.

the end of the section so as to deal with non-interactive sums (we omit details of the straightforward computations).

Before proceeding, we wish to point out a remarkable case of *irrelevance*, cf. Sect. 2. In $Z^+$ one sums up $2n$ terms: of course these terms are *interactive*, because each $X_i$ appears twice. Let us "force" non-interactivity: for each $i$ we take $Y_i$ equidistributed with $X_i^+$; the whole $2n$-tuple $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ is assumed to be non-interactive, by this precluding that contributions relative to distinct $i$'s might interact.

**Proposition 1** *The fuzzy numbers $Z^+$ and $\tilde{Z}^+ \doteq \sum_i X_i^2 + 2\sum_i |s_i| Y_i$ are equidistributed.*

*Proof* Fix $i$; we use Lemma 1 with $x \circ y \doteq x^2 + 2|s_i|y$, a non-commutative operation which, as required in the lemma, is order-preserving on the corresponding support. This proves the proposition for $n = 1$; if $n > 1$ just observe that terms relative to distinct $i$'s are non-interactive.

From now on, we shall assume $\lim_n s^{(n)} = s$, $s \in \mathbb{R}$, a fact which will turn out to hold in the case of Gaussian channels. In the sequel we rather need $W_n \doteq \frac{Z}{n}$; we shall present directly its *asymptotic* form $W$, obtained when $n$ diverges and so $s^{(n)}$ tends to the constant $s$:

$$W(w) = 1 + s - \sqrt{w + s^2}\ , \quad w \in [0, 1 + 2s]$$

The total area $\int_o^{1+2s} W(w)dw$ is readily computed to be $s + \frac{1}{3}$, which is so the *total fuzziness* of the fuzzy quantity $W$.

We are now ready to introduce the allowed *fuzziness error*, which sides with the allowed *probabilistic error*, as currently used in "crisp" information theory. First some heuristics: assume $\eta \le d$, both crisp numbers. Say $B$ is triangular $(a, b, c)$, or "triangular-like", with $\eta$ ($\eta < c$) so "near" to the extreme $c$ that you are willing to consider "negligible" the area, i.e. the fuzziness, corresponding to $[\eta, c]$: in this case, with a *fuzziness error* = *negligible area*, you might accept to keep the inequality $B \le d$ which would still be "roughly" true, i.e. it would be true up to the allowed fuzziness error. Following this philosophy, we fix the allowed fuzziness error $F \in [0, s + \frac{1}{3}[$ and search for the $\xi_F = \xi_F(s)$ such that the area corresponding to $[\xi_F, 1 + 2s]$ is equal to $F$, and is therefore "negligible".

**Definition 1** Let $\xi_F = \xi_F(s)$ be the unique solution of the equation

$$\int_{\xi_F}^{1+2s} W(w)dw = F \tag{4}$$

To compute $\xi_F$ one has to solve the integral, which by linear transformations can be taken back to the standard integral $\sqrt{u}\,du$; it turns out that, to obtain now $\xi_F$, one has to solve a cubic equation in $v = \sqrt{u}$. This can be done by using Cardano's formula, which leads to quite an unwieldy expression. We will be contented to know that in principle $\xi_F$ can be computed, even if not without effort. To proceed, we find it convenient to resort also to an upper bound $\zeta_F > \xi_F$. We replace the *convex* distribution function $W(w)$ by the linear[4] distribution function $W^*(w) = 1 - \frac{w}{1+2s} \geq W(w)$; the total fuzziness of $W^*$ goes up to $s + \frac{1}{2}$. Let us solve the problem as in definition 1, only with $\zeta_F$ and $W^*$ instead of $\xi_F$ and $W$. One soon obtains:

$$\zeta_F = \zeta_F(s) = 1 + 2s - \sqrt{2(1+2s)F} \;\; > \;\; \xi_F \tag{5}$$

*Addendum*: We shortly cover the familiar case of a *non-interactive* sum $Z = X + Y$, $f_{XY}(x, y) = X(x) \wedge Y(y)$, when $X(x)$ and $Y(y)$ are continuous and strictly increasing on $[0, a]$ and $[0, b]$, respectively; we assume without real restriction $a \leq b$; $Y(z - x)$ is strictly decreasing when seen as a functions of $x$. For fixed $z$ in the interval $[0, a + b]$, one soon checks that the equation in $x$ $X(x) = Y(z - x)$ has a single solution $\mu(z)$, $X\big(\mu(z)\big) = Y\big(z - \mu(z)\big)$, and that $\mu(z)$ strictly increases in $z$ from $\mu(0) = 0$ to $\mu(a + b) = 1$. More specifically, one conveniently distinguishes three cases, $0 \leq z \leq a$, $a \leq z \leq b$ (void for $a = b$) and $b \leq z \leq a + b$. In the first, $x = z - y \in [0, a] \cap [z - b, z] = [0, z]$; for fixed $z \in [0, a]$, on the border $x = 0$ the increasing function $X(x)$ and the decreasing function $Y(z - x)$ take the two values $0 = X(0) < Y(z)$, while on the border $x = z$ they take the two values $X(z) > Y(0) = 0$: the required maximum is found for $x = \mu(z)$. The remaining two cases are dealt in the same way and give the same solution $Z(z) = X\big(\mu(z)\big) = Y\big(z - \mu(z)\big)$, which is found also when $X(x)$ and $Y(y)$ are both strictly decreasing.

## 4 Channel Capacity Under a Fuzziness Constraint

Channel capacity, in particular the capacity of a Gaussian channel, is an *asymptotic* notion, being the limit of rates of optimal codes as the codeword length $n$ diverges; roughly speaking, it is the maximal speed at which data can be *reliably* sent over the noisy channel, the channel decoder working with a "negligible" error probability. We refer the reader e.g. to [5], where, for given *energy power* $\Pi$, e.g. $\Pi = P$, one proves the famous Shannon theorem on the capacity $C$ of a Gaussian channel by using a random coding technique, as did Shannon himself:

$$C = \frac{1}{2}\log_2\big(1 + \frac{\Pi}{N}\big)$$

where the capacity is measured in *bits* and $N$ is the variance of the white noise added to each crisp signal. If the code used is *optima*l for given power $\Pi$ and $s_1, \ldots, s_n$ is

[4] This "forced linearization" is not new in soft computing, when one replaces genuine products of triangular numbers by *pseudo-products* which "force" linearity on the result; in practice, going to $W^*$ amounts to replace genuine squares as in (3) by *pseudo-squares*.

one of its codewords, one can assume[5] that, as the codeword length increases, the sum of squares $\frac{1}{n}\sum s_i^2$ tends to $\Pi$ (so the energy power bound is asymptotically verified with equality), while $\frac{1}{n}\sum |s_i|$ tends to $s = \sqrt{\frac{2}{\pi}\Pi}$, which is the value of $s$ to be inserted in Problem 1, Sect. 3, and in (5) to obtain $\xi_F$ and $\zeta_F$ directly as functions of $\Pi$, rather than $s$, $\xi_F = \xi_{F,\Pi}$ and $\zeta_F = \zeta_{F,\Pi}$.

Let us explore the consequences of fuzziness up to tolerated fuzziness $F$. Unfortunately, we do not control crisply each signal $s_i$ due e.g. to instrumental imprecision, and so the crisp signal actually fed to the channel might not be $s_i$. Let us construct the optimal code assuming energy power $\Pi$. In the *worst case* the signal power might actually be as high as $\Pi + \xi_{F,\Pi}$, cf. Definition 1, Sect. 3, and recall that the sum of squares tends to $\Pi$. We shall pessimistically assume that this worst-case situation does actually occur. With given power $P$, an *ad hoc* way out of the snag might be to construct the optimal code with respect to the diminished energy power $P - \xi_{F,P}$. Actually this difference might be $\leq 0$, in which case the code meant to fight worst case fuzziness *cannot* be constructed: to signal this fact one has to write $|P - \xi_{F,P}|^+$ instead of $P - \xi_{F,P}$, where $x^+ = x$ for $x$ positive, else is 0. E.g. in the limit case $F = 0$, when $\xi_{0,P} = \zeta_{0,P} = 1 + 2s = 1 + 2\sqrt{\frac{2}{\pi}P}$, the construction is possible only for $P > 1 + 2\sqrt{\frac{2}{\pi}P}$. Solving[6] the inequality:

$$P > 1 + \frac{4 + 2\sqrt{2\pi + 4}}{\pi} \approx 4.31$$

Actually, to determine the energy power $\Pi \leq P$ according to which the optimal code has to be constructed, one should rather proceed as follows. One has $\Pi = P - \xi_{F,\Pi}$ (if this difference is not positive there is no possibility to ensure worst-case reliable transmission), while $\xi_{F,\Pi}$ is obtained as in (4) by Cardano's formula, $s = \sqrt{\frac{2}{\pi}\Pi}$. This allows one to obtain $\Pi$ and $\xi_F$ as functions of $P$ and $F$, $\Pi = \Pi_{F,P}$, $\xi_F = \xi_{F,\Pi} \doteq \rho_{F,P}$. To understand what happens, if one is contented with the lower bound $\zeta_F$ rather than $\xi_F$, cf. footnote 4, the equations to solve, cf. (5), would become:

$$P - \zeta = \Pi \quad \text{and} \quad \zeta = 1 + 2\sqrt{\frac{2\Pi}{\pi}} - \sqrt{2\left(1 + 2\sqrt{\frac{2\Pi}{\pi}}\right)F}$$

[5] Just as a hint, using the random coding technique to construct optimal codes, in a codeword $s_1 \dots s_n$, each $s_i$ can be seen as the output of a Gaussian distribution $\mathcal{N}(0, \Pi)$, where the expected normalized sum of squares of the random outputs is constrained to be $\Pi$; expectation is unconditionally additive, and so the expectation of the normalized sum of the absolute values is soon computed to be $\sqrt{(2/\pi)\Pi}$, a value approximated with high probability by the actual outputs if $n$ is large; cf. [4] for details.

[6] Note that we are assuming triangular fuzzy signals of unit fuzziness: were it not so, the last bound would be on $\frac{P}{c^2}$ rather than $P$, cf. footnote 2.

in the two unknowns[7] $\Pi$ and $\zeta$. In this paper we shall not deal with the taxing numerical task of approximating to the desired degree of precision $\Pi_{F,P}$ and $\rho_{F,P}$ for given values of $F$ and $P$; however, the two parameters are *well defined*, if only "in principle".

**Theorem 2** *To ensure worst-case reliable transmission with triangular fuzzy input signals of unit fuzziness bound to verify the fuzzy energy power constraint (3) with tolerated fuzziness error $F$, the capacity of the Gaussian channel is*

$$C = \frac{1}{2}\log_2\left(1 + \frac{|P - \rho_{F,P}|^+}{N}\right)$$

## References

1. Bortolussi L, Sgarro A (2010) Possibilistic coding: error detection vs. error correction. In: Borgelt Ch et al (ed) Combining soft computing and statistical methods in data analysis. Advances in intelligent and soft computing, vol 77. Springer, pp 41–48
2. Bortolussi L, Dinu LP, Franzoi L, Sgarro A (2015) Coding theory: a general framework and two inverse problems. Fundam Inf 141:297–310
3. Carlsson C, Fullér R, Majlender P (2004) Additions of completely correlated fuzzy numbers. In: IEEE international conference on fuzzy systems, pp 535–539
4. Coroianu L, Fullér R (2013) On multiplication of interactive fuzzy numbers. In: IEEE International symposium on intelligent systems and informatics, SISY, pp 181–185
5. Cover TM, Thomas JA (2006) Elements of information theory. Willey series in telecommunications. Wiley, New York
6. Dubois D, Prade H (1981) Additions of interactive fuzzy numbers. IEEE Trans Autom Control AC-26 4:926–930
7. Dubois D, Prade H (2000) Fundamentals of fuzzy sets. Kluwer Academic Publishers
8. Franzoi L, Sgarro A (2015) (Ir)relevance of interactivity in fuzzy arithmetic. Math Pannonica 25(1):93–103
9. Fullér R, Majlender P (2004) On interactive fuzzy numbers. Fuzzy Sets Syst 143(3):353–369
10. Klir GJ, Folger TA (1988) Fuzzy Sets. Uncertainty and information. Prentice Hall
11. Sgarro A (2002) Possibilistic information theory: a coding-theoretic approach. Fuzzy Sets Syst 132(2):11–32
12. Sgarro A, Franzoi L (2012) Fuzzy arithmetic for fuzzy n-poles. In: Proceedings of IPMU, vol 3, pp 1–8
13. Sgarro A, Franzoi L (2016) (Ir)relevant T-norm joint distributions in the arithmetic of fuzzy quantities. Accepted by IPMU 2016, Eindhoven, Netherlands, 20–24 June 2016

[7] As a hint to computations, from the second equation obtain $u \doteq 1 + 2\sqrt{(2/\pi)\Pi}$ as a function of $\zeta$ and $F$ to be replaced in the first equation re-written as $1 + 2\sqrt{(2/\pi)(P-\zeta)} = u$; for given $F$ and $P$ deal numerically with the resulting fourth-degree equation in $\zeta$.

# Fuzzy Clustering Through Robust Factor Analyzers

**Luis Angel García-Escudero, Francesca Greselin and Agustin Mayo Iscar**

**Abstract** In fuzzy clustering, data elements can belong to more than one cluster, and membership levels are associated with each element, to indicate the strength of the association between that data element and a particular cluster. Unfortunately, fuzzy clustering is not robust, while in real applications the data is contaminated by outliers and noise, and the assumed underlying Gaussian distributions could be unrealistic. Here we propose a robust fuzzy estimator for clustering through Factor Analyzers, by introducing the joint usage of trimming and of constrained estimation of noise matrices in the classic Maximum Likelihood approach.

## 1 Introduction

Clustering can be considered the most important unsupervised learning problem. It is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called clusters. A cluster is therefore a collection of objects which are similar to one another and thus can be treated collectively as one group. Clustering algorithms may be classified into Exclusive (or Crisp, Hard), Overlapping, Hierarchical and Probabilistic. To recall some well known examples, K-means [12] is an exclusive clustering algorithm, Fuzzy C-means [2] is an overlapping clustering algorithm, Single-linkage [1] is an agglomerative hierarchical clustering and, lastly, Mixture of Gaussian is a probabilistic clustering algorithm. In the present work, we move from

L.A. García-Escudero · A. Mayo Iscar
Department of Statistics and Operational Research and IMUVA,
University of Valladolid, Valladolid, Spain
e-mail: lagarcia@eio.uva.es

A. Mayo Iscar
e-mail: agustin@med.uva.es

F. Greselin (✉)
Department of Statistics and Quantitative Methods,
Milano-Bicocca University, Milan, Italy
e-mail: francesca.greselin@unimib.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_29

a robust constrained fuzzy clustering approach based on Gaussian components [3], and we introduce a fuzzy version of Mixtures of Gaussian Factor Analyzers (MFA).

Starting from Wee and Fu's seminal work [16], fuzzy clustering has received an increasing attention by researchers from several fields in the last fifty years. The aim is to discover a limited number of homogeneous clusters in such a way that the objects are assigned to the clusters according to the so-called membership degrees ranging in the interval [0, 1]. In real applications, the data is bound to have noise and outliers, and the assumed models such as Gaussian distributions are only approximations to reality. Unfortunately, one of the main limitations of all clustering algorithms is that they are not robust to noise: a small fraction of outlying data may drastically deteriorate the clustering ability. Hence we will provide robustness properties to our estimator for Gaussian Factor Analyzers, by trimming those observations that are less plausible under the estimated model. According to [10], a robust procedure can be characterized by the following: (1) it should have a reasonably good efficiency (accuracy) at the assumed model; (2) small deviations from the model assumptions should impair the performance only by a small amount; and (3) larger deviations from the model assumptions should not cause a catastrophe. We could see that our proposal satisfies the three properties.

## 2 Fuzzy Clustering Through Gaussian Factors

Suppose that we have $n$ observations $\{\mathbf{x}_1 \ldots \mathbf{x}_n\}$ in $\mathbb{R}^p$ and we want to fuzzy-classify them into $k$ clusters. Therefore, our aim is to obtain a collection of non-negative membership values $u_{ij} \in [0, 1]$ for all $i = 1 \ldots n$ and $j = 1 \ldots k$. Increasing degrees of membership are allowed when $u_{ij} \in (0, 1)$, while $u_{ij} = 1$ indicates that object $i$ fully belongs to cluster $j$ and, conversely, $u_{ij} = 0$ means that it does not belong to this cluster. We will denote an observation as fully trimmed if $u_{ij} = 0$ for all $j = 1 \ldots k$ and, thus, this observation has no membership contribution to any cluster.

Further, we want to employ Factor Analysis and suppose that, as in many phenomena, the $p$ observed variables could be explained by a few unobserved ones. Factor Analysis is an effective method of summarizing the variability between a number of correlated features, through a much smaller number of unobservable, hence named *latent*, factors. Under this approach, each single variable (among the $p$ observed ones) is assumed to be a linear combination of $d$ underlying common factors with an accompanying error term to account for the part of the variability which is unique to it (not in common with other variables). We will assume that the distribution of $\mathbf{x}_i$ can be given as

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U}_i + \mathbf{e}_i \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\Lambda}$ is a $p \times d$ matrix of *factor loadings*, the *factors* $\mathbf{U}_1, \ldots, \mathbf{U}_n$ are $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ distributed independently of the *errors* $\mathbf{e}_i$. The latter are independently $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ distributed, and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix. The diagonality of $\boldsymbol{\Psi}$ is one of the key assumptions of factor analysis: the observed variables are independent given the factors. Note that the factor variable $\mathbf{U}_i$ models correlations between the elements of $\mathbf{x}_i$, while the errors $\mathbf{e}_i$ account for independent noise for $\mathbf{x}_i$. We suppose that $d < p$. Under these assumptions, $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ has the form

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{2}$$

Given a fixed trimming proportion $\alpha \in [0, 1)$, a fixed constant $c \geq 1$ and a fixed value of the fuzzifier parameter $m > 1$, a robust constrained fuzzy clustering problem can be defined through the maximization of the objective function

$$\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m} \log \phi(\mathbf{x}_i; \mathbf{m}_j, \mathbf{S}_j), \tag{3}$$

where $\phi(\cdot; \mathbf{m}, \mathbf{S})$ is the density of the multivariate Gaussian with mean $\mathbf{m}$ and covariance $\mathbf{S}$, and the membership values $u_{ij} \geq 0$ are assumed to satisfy

$$\sum_{j=1}^{k} u_{ij} = 1 \quad \text{if} \quad i \in \mathcal{I} \quad \text{and} \quad \sum_{j=1}^{k} u_{ij} = 0 \quad \text{otherwise}, \tag{4}$$

for a subset

$$\mathcal{I} \subset 1, 2, \ldots, n \quad \text{with} \quad \#\mathcal{I} = [n(1-\alpha)], \tag{5}$$

where $\mathbf{m}_1, \ldots, \mathbf{m}_k$ are vectors in $\mathbb{R}^p$, and $\mathbf{S}_1, \ldots, \mathbf{S}_k$ are positive semidefinite $p \times p$ matrices satisfying the decomposition in (2), i.e. $\mathbf{S}_j = \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j' + \boldsymbol{\Psi}_j$. With reference to the diagonal elements $\{\psi_k\}_{k=1,\ldots,p}$ of the noise matrices $\boldsymbol{\Psi}_j$, it is required that

$$\psi_{j_1 h} \leq c_{noise} \;\; \psi_{j_2 l} \qquad \text{for every } 1 \leq h \neq l \leq p \text{ and } 1 \leq j_1 \neq j_2. \leq k \tag{6}$$

The constant $c_{noise}$ is finite and such that $c_{noise} \geq 1$, to avoid the $|S_j| \to 0$ case. This constraint can be seen as an adaptation to MFA of those introduced in [5, 11], and is similar to the mild restrictions implemented for MFA in [7]. They all go back to the seminal paper [9].

Notice that $u_{i1} = \ldots = u_{ik} = 0$ for all $i \notin \mathcal{I}$, so the observations in $\mathcal{I}$ do not contribute to the summation in the target function (3).

Our fuzzy method is based on a maximum likelihood criterium defined on a specific underlying statistical model, as in many other proposal in the literature.

After the introduction of trimmed observation, the second specific features of the proposed methodology is the application of the eigenvalue ratio constraint in (6). This is needed to avoid the unboundedness of the objective function (3), whenever one of the $\mathbf{m}_j$ is equal to one of the observations $\mathbf{x}_i$, setting $u_{ij} = 1$, and for a sequence of scatter matrices $\mathbf{S}_j$ such that $|\mathbf{S}_j| \to 0$. This problem is recurrent in Cluster Analysis whenever general scatter matrices are allowed, and has been already noticed in fuzzy clustering, among other authors, by [8]. In our approach, the unboundedness problem is addressed by constraining the ratio between the largest and smallest eigenvalues of the so-called noise matrices $\boldsymbol{\Psi}_j$. Larger values of $c_{noise}$ lead to an almost unconstrained fuzzy clustering approach.

It is well known that the use of an objective function like that in (3) tends to provide clusters with similar sizes, or more precisely, with similar values of $\sum_{i=1}^{n} u_{ij}^{m}$. If this effect is not desired then it is better to replace the objective function (3) by

$$\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m} \log p_j \phi(\mathbf{x}_i; \mathbf{m}_j, \mathbf{S}_j), \tag{7}$$

where $p_j \in [0, 1]$ and $\sum_{j=1}^{k} p_j = 1$ are some weights to be maximized in the objective function, as in the entropy regularizations in [14]. Once the membership values are known, the weights are optimally determined as

$$p_j = \sum_{i=1}^{n} u_{ij}^{m} / \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m}$$

(see [4], for a detailed explanation). Finally, considering (7) as our target function, and performing trimming and constrained estimation along the EM algorithm we obtain a robust approach to fuzzy clustering through factor analyzers.

More precisely, we consider an AECM algorithm, where we incorporates a concentration step, as in many high-breakdown point robust algorithms like [15], before each E-step. After selecting the set of observations that contributed the most to the target function (concentration step), at each iteration, given the values of the parameters, the best possible membership values are obtained (E-step). Afterwards, the parameters are updated by maximizing expression (7) on the parameters (M-step). The name of AECM (that appeared in the literature for the case of mixtures of Gaussian factor analyzers, see [13]) comes from the fact that the M-step is performed alternatively on a partition of the parameter space. When updating the $\mathbf{S}_j$ matrices the constraint on the eigenvalue ratios are imposed accordingly, along the lines of [3].

Finally, it is worth to remark that the general approach presented herein encompasses the soft robust clustering method introduced in [6], and leads to hard clustering for $m = 1$. For $m > 1$ it provides fuzzy clustering.

## 3 Numerical Results

We present here a first experiment on synthetic data, to show the performance of the proposal. We choose a two component population in $\mathbb{R}^{10}$, from which we draw two samples. Aiming at providing a plot of the obtained results, we work with unidimensional factors (otherwise we could not find a unique space, for the two components, to represent the data). The first population $X_1$ is defined as follows:

$$X_{11} \sim \mathcal{N}(0,1) + 4 \qquad X_{12} \sim 5 * X_{11} + 3 * \mathcal{N}(0,1) - 6;$$

and the second population $X_2$ is given as:

$$X_{21} \sim \mathcal{N}(0,1) + 4 \qquad X_{22} \sim X_{21} + 2 * \mathcal{N}(0,1) + 19.$$

After drawing 100 points for each component, to check the robustness of our approach, we add some pointwise contamination $X_3$ to the data, by drawing 10 points as follows

$$X_{31} \sim \mathcal{N}(0,1) + 4 \qquad X_{32} \sim 50 + 0.01 * \mathcal{N}(0,1);$$

and 10 more points, denoted by $X_4$, where

$$X_{41} \sim \mathcal{N}(0,1) + 6 \qquad X_{42} \sim -20 + 0.01 * \mathcal{N}(0,1).$$

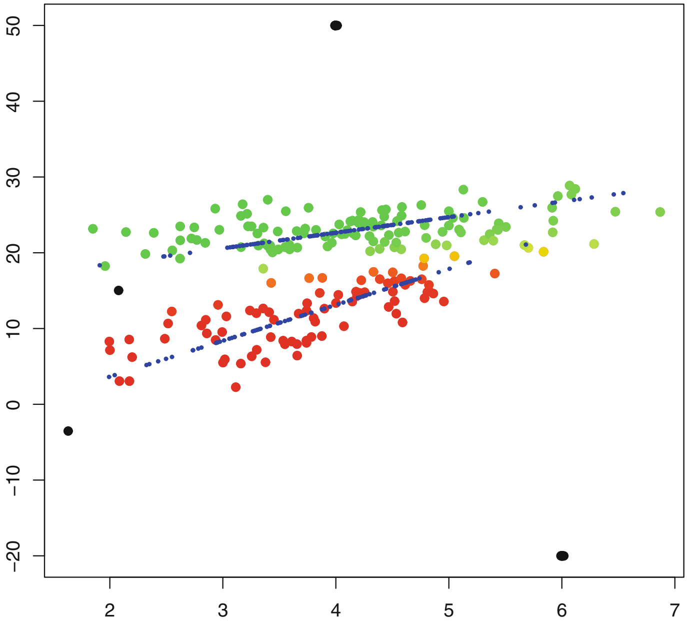


**Fig. 1** Fuzzy classification of the synthetic data. *Blue* points are the projections of the 10-dimensional data in the latent factor space of the first component. *Black* points are trimmed units. The strength of the membership values is represented by the color usage

Finally, we complement the data matrix with $X_{ij} \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, 4$ and $j = 3, \ldots, 10$. In this way we have built a dataset where one factor is explaining the correlation among the 10 variables, in each component.

Figure 1 shows that the estimation is robust to the most dangerous outliers, in the form of pointwise contamination (although we have used the ten variables when applying the algorithm, only the first two variables are represented here).

## 4 Concluding Remarks

We have introduced Fuzzy and robust estimation of mixtures of Factor Analyzers, by including a trimming procedure and constrained evaluation of the noise matrices along the steps of the EM algorithm. Our proposal lays in between soft and hard robust clustering, and encompasses them. Based on our first findings, we observed that small deviations from the model assumptions impair the performance of the fuzzy classifier only by a small amount, and that good efficiency is obtained on data without contamination. Further work is needed to show the advantages of the proposed approach in real data applications.

## References

1. Cattell R (1944) A note on correlation clusters and cluster search methods. Psychometrika 9(3):169–184
2. Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybern 3:32–57
3. Fritz H, García-Escudero LA, Mayo-Iscar A (2013) A fast algorithm for robust constrained clustering. Comput Stat Data Anal 61:124–136
4. Fritz H, García-Escudero LA, Mayo-Iscar A (2013) Robust constrained fuzzy clustering. Inf Sci 245:38–52
5. García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. Ann Stat 36(3):1324–1345
6. García-Escudero LA, Gordaliza A, Greselin F, Ingrassia S, Mayo-Iscar A (2016) The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. Comput Stat Data Anal 99:131–147
7. Greselin F, Ingrassia S (2015) Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers. Stat Comput 25:215–226
8. Gustafson EE, Kessel WC (1979) Fuzzy clustering with a fuzzy covariance matrix. In: Proceedings of the IEEE lnternational conference on fuzzy systems, San Diego, pp 761–766
9. Hathaway R (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Stat 13(2):795–800
10. Huber PJ (1981) Robust statistics. Wiley, New York
11. Ingrassia S, Rocci R (2007) Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. Comput Stat Data Anal 51:5339–5351
12. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 281–297

13. McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
14. Miyamoto S, Mukaidono M (1997) Fuzzy c-means as a regularization and maximum entropy approach. In: Proceedings of the 7th international fuzzy systems association world congress (IFSA97), vol 2, pp 86–92
15. Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
16. Wee WG, Fu KS (1969) A formulation of fuzzy automata and its application as a model of learning systems. IEEE Trans Syst Sci Cybern 5(3):215–223. doi:10.1109/TSSC.1969.300263

# Consensus-Based Clustering in Numerical Decision-Making

José Luis García-Lapresta and David Pérez-Román

**Abstract** In this paper, we consider that a set of agents assess a set of alternatives through numbers in the unit interval. In this setting, we introduce a measure that assigns a degree of consensus to each subset of agents with respect to every subset of alternatives. This consensus measure is defined as 1 minus the outcome generated by a symmetric aggregation function to the distances between the corresponding individual assessments. We establish some properties of the consensus measure, some of them depending on the used aggregation function. We also introduce an agglomerative hierarchical clustering procedure that is generated by similarity functions based on the previous consensus measures.

## 1 Introduction

When a group of agents show their opinions abut a set of alternatives, an important issue is to know the homogeneity of these opinions. In this paper we consider that agents evaluate each alternative by means of a number in the unit interval. For measuring the consensus in a group of agents over a subset of alternatives, we propose to aggregate the distances between the corresponding individual assessments through an appropriate symmetric aggregation function. This outcome measures the dispersion of individual opinions in a similar way to the Gini index [13] measures the inequality of individual incomes.

The consensus measure we propose is just 1 minus the mentioned dispersion measure. The most important is not to know the degree of consensus in a specific group of agents, but comparing the consensus of different groups of agents with

J.L. García-Lapresta (✉)
PRESAD Research Group, BORDA Research Unit, IMUVA,
Dept. de Economía Aplicada, Universidad de Valladolid, Valladolid, Spain
e-mail: lapresta@eco.uva.es

D. Pérez-Román
PRESAD Research Group, BORDA Research Unit, Dep. de Organización de Empresas
y C.I.M., Universidad de Valladolid, Valladolid, Spain
e-mail: david@emp.uva.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_30

respect to an alternative or a subset of alternatives. This is the starting point of the agglomerative hierarchical clustering procedure we propose. We consider as linkage clustering criterion one generated by a consensus-based similarity function that merges clusters or individuals by maximizing the consensus.

The rest of the paper is organized as follows. Section 2 includes some notation and basic notions. In Sect. 3 we include our proposal for measuring the consensus. Section 4 is devoted to introduce the agglomerative hierarchical clustering procedure. And Sect. 5 concludes with some remarks and further research.

## 2 Preliminaries

Along the paper, vectors in $[0,1]^k$ are denoted as $\boldsymbol{y} = (y_1, \ldots, y_k)$; in particular, $\mathbf{0} = (0, \ldots, 0)$ and $\mathbf{1} = (1, \ldots, 1)$. Given $\boldsymbol{y}, \boldsymbol{z} \in [0,1]^k$, by $\boldsymbol{y} \geq \boldsymbol{z}$ we mean $y_i \geq z_i$ for every $i \in \{1, \ldots, k\}$. With $\#I$ we denote the cardinality of $I$. With $\mathcal{P}_2(A) = \{I \subseteq A \mid \#I \geq 2\}$ we denote the family of subsets of at least two agents.

We begin by defining standard properties of real functions on $[0,1]^k$ and aggregation functions. For further details the interested reader is referred to Beliakov et al. [5], Grabisch et al. [14] and Beliakov et al. [4].

**Definition 1**

1. Given $k \in \mathbb{N}$, a function $F^{(k)} : [0,1]^k \longrightarrow [0,1]$ is *symmetric* if for all permutation $\pi$ on $\{1, \ldots, k\}$ and $\boldsymbol{y} \in [0,1]^k$ it holds that $F^{(k)}(y_{\pi(1)}, \ldots, y_{\pi(k)}) = F^{(k)}(y_1, \ldots, y_k)$.
2. Given $k \in \mathbb{N}$, a function $F^{(k)} : [0,1]^k \longrightarrow [0,1]$ is *monotonic* if for all $\boldsymbol{y}, \boldsymbol{z} \in [0,1]^k$ it holds that $\boldsymbol{y} \geq \boldsymbol{z} \;\Rightarrow\; F^{(k)}(\boldsymbol{y}) \geq F^{(k)}(\boldsymbol{z})$.
3. Given $k \in \mathbb{N}$, a function $F^{(k)} : [0,1]^k \longrightarrow [0,1]$ is called an *k-ary aggregation function* if it is monotonic and satisfies the boundary conditions $F^{(k)}(\mathbf{0}) = 0$ and $F^{(k)}(\mathbf{1}) = 1$. In the extreme case $k = 1$, the convention $F^{(1)}(y) = y$ for every $y \in [0,1]$ is considered.
4. An *aggregation function* is a sequence $F = \left(F^{(k)}\right)_{k \in \mathbb{N}}$ of $k$-ary aggregation functions.
5. An aggregation function $F = \left(F^{(k)}\right)_{k \in \mathbb{N}}$ is symmetric (monotonic) whenever $F^{(k)}$ is symmetric (monotonic) for every $k \in \mathbb{N}$.

For the sake of simplicity, the $k$-arity is omitted whenever it is clear from the context.

## 3 Consensus

For measuring the degree of consensus among a group of agents that provide their opinions on a set of alternatives, different proposals can be found in the literature (see Martínez-Panero [16] for an overview of different notions of consensus).

In the social choice framework, the notion of *consensus measure* was introduced by Bosch [6] in the context of linear orders. Additionally, Bosch [6] and Alcalde-Unzu and Vorsatz [2] provided axiomatic characterizations of several consensus measures in the context of linear orders. García-Lapresta and Pérez-Román [9] extended that notion to the context of weak orders and they analyzed a class of consensus measures generated by distances. Alcantud et al. [3] provided axiomatic characterizations of some consensus measures in the setting of approval voting. In turn, Erdamar et al. [7] extended the notion of consensus measure to the preference-approval setting through different kinds of distances, and García-Lapresta et al. [12] introduced another extension to the framework of hesitant linguistic assessments.

Let $A = \{1, \ldots, m\}$, with $m \geq 2$, be a set of agents and let $X = \{x_1, \ldots, x_n\}$, with $n \geq 2$, be the set of alternatives which have to be evaluated in the unit interval.

A *profile* is a matrix

$$V = \begin{pmatrix} v_1^1 & \cdots & v_i^1 & \cdots & v_n^1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ v_1^a & \cdots & v_i^a & \cdots & v_n^a \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ v_1^m & \cdots & v_i^m & \cdots & v_n^m \end{pmatrix} = \left(v_i^a\right)$$

consisting of $m$ rows and $n$ columns of numbers in [0, 1], where the element $v_i^a$ represents the assessment given by the agent $a \in A$ to the alternative $x_i \in X$.

Let $V = \left(v_i^a\right)$ be a profile, $\pi$ a permutation on $A$, $\sigma$ a permutation on $\{1, \ldots, n\}$, $I \in \mathcal{P}_2(A)$ and $\emptyset \neq Y \subseteq X$. The profiles $V^\pi$, $V_\sigma$ and $V^{-1}$, and the subsets $I^\pi$ and $Y_\sigma$ are defined as follows:

1. $V^\pi = \left(u_i^a\right)$ where $u_i^a = v_i^{\pi(a)}$.
2. $V_\sigma = \left(u_i^a\right)$ where $u_i^a = v_{\sigma(i)}^a$.
3. $V^{-1} = \left(u_i^a\right)$ where $u_i^a = 1 - v_i^a$.
4. $I^\pi = \left\{\pi^{-1}(a) \mid a \in A\right\}$, i.e., $a \in I^\pi \Leftrightarrow \pi(a) \in I$.
5. $Y_\sigma = \{x_{\sigma^{-1}(i)} \mid x_i \in Y\}$, i.e., $x_i \in Y_\sigma \Leftrightarrow x_{\sigma(i)} \in Y$.

**Definition 2** Let $F = \left(F^{(k)}\right)_{k \in \mathbb{N}}$ be a symmetric aggregation function. Given a profile $V = (v_i^a)$, the *degree of consensus* in a subset of agents $I \in \mathcal{P}_2(A)$ over a subset of alternatives $\emptyset \neq Y \subseteq X$ is defined as

$$C_F(V, I, Y) = 1 - F\left(\left|v_i^a - v_i^b\right|_{\substack{a,\, b \in I,\, a<b \\ x_i \in Y}}\right).$$

In Proposition 1 we establish some properties of the consensus notion introduced in Definition 2. Normalization means that the degree of consensus is always in the unit interval. Anonymity means that all agents are treated in the same way. Unanimity establishes necessary and sufficient conditions for reaching maximum consensus. Maximum dissension establishes necessary and sufficient conditions for reaching minimum consensus in two agents. Positiveness establishes that with more than two agents the degree of consensus is never minimum. Neutrality means that all alternatives are treated in the same way. And reciprocity means that if all the agents reverse their assessments, then the degree of consensus does not change.

**Proposition 1** *Let $F = \big(F^{(k)}\big)_{k\in\mathbb{N}}$ be an aggregation function. The following properties are satisfied:*

1. Normalization*: $C_F(V, I, Y) \in [0, 1]$.*
2. Anonymity*: $C_F(V^\pi, I^\pi, Y) = C_F(V, I, Y)$ for every permutation $\pi$ on $A$.*
3. Unanimity*: If for every $x_i \in Y$ there exists $t_i \in [0, 1]$ such that $v_i^a = t_i$ for every $a \in I$, then $C_F(V, I, Y) = 1$.*
   *Additionally, if $F^{(k)}(\mathbf{y}) = 0 \Leftrightarrow \mathbf{y} = \mathbf{0}$, for all $k \in \mathbb{N}$ and $\mathbf{y} \in [0, 1]^k$, and $C_F(V, I, Y) = 1$, then for every $x_i \in Y$ there exists $t_i \in [0, 1]$ such that $v_i^a = t_i$ for every $a \in I$.*
4. Maximum dissension*: If $\Big(\big(v_i^a = 0 \text{ and } v_i^b = 1\big) \text{ or } \big(v_i^a = 1 \text{ and } v_i^b = 0\big)\Big)$ for all $x_i \in Y$, then $C_F(V, \{a, b\}, Y) = 0$.*
   *Additionally, if $F^{(k)}(\mathbf{y}) = 1 \Leftrightarrow \mathbf{y} = \mathbf{1}$, for all $k \in \mathbb{N}$ and $\mathbf{y} \in [0, 1]^k$, and $C_F(V, \{a, b\}, Y) = 0$, then $\Big(\big(v_i^a = 0 \text{ and } v_i^b = 1\big) \text{ or } \big(v_i^a = 1 \text{ and } v_i^b = 0\big)\Big)$ for all $x_i \in Y$.*
5. Positiveness*: If $F^{(k)}(\mathbf{y}) = 1 \Leftrightarrow \mathbf{y} = \mathbf{1}$, for all $k \in \mathbb{N}$ and $\mathbf{y} \in [0, 1]^k$, and $\#I > 2$, then $C_F(V, I, Y) > 0$.*
6. Neutrality*: $C_F(V_\sigma, I, Y_\sigma) = C_F(V, I, Y)$ for every permutation $\sigma$ on $\{1, \ldots, n\}$.*
7. Reciprocity*: $C_F(V^{-1}, I, Y) = C_F(V, I, Y)$.*

*Proof* It is straightforward.

## 4 Clustering

There are many clustering algorithms (see Ward [17], Jain et al. [15] and Everitt et al. [8], among others). Most methods of hierarchical clustering use an appropriate metric (for measuring the distance between pairs of observations), and a linkage criterion which specifies the similarity/dissimilarity of sets as a function of the pairwise distances of observations in the corresponding sets.

Ward [17] proposed an agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimization of an objective function. In the following procedure, the criterion is to maximize the consensus.

**Definition 3** Let $F = \big(F^{(k)}\big)_{k\in\mathbb{N}}$ be an aggregation function. Given a profile $V = (v_i^a)$, the *similarity function* relative to a subset of alternatives $\emptyset \neq Y \subseteq X$

$$S_F^Y : \big(\mathcal{P}(A) \setminus \{\emptyset\}\big)^2 \longrightarrow [0, 1]$$

is defined as

$$S_F^Y(I,J) = \begin{cases} C_F(V, I \cup J, Y), & \text{if } \#(I \cup J) \geq 2, \\ 1, & \text{if } \#(I \cup J) = 1. \end{cases}$$

*Remark 1* In the extreme case of two agents and a single alternative, the similarity between these agents on that alternative is just 1 minus the distance between their assessments: given an alternative $x_i \in X$ and two different agents $a, b \in A$, we have

$$S_F^{\{x_i\}}\left(\{a\},\{b\}\right) = C_F\left(V, \{a,b\}, \{x_i\}\right) = 1 - \left|v_i^a - v_i^b\right|.$$

The agglomerative hierarchical clustering procedure we propose is related to the ones provided by García-Lapresta and Pérez-Román [10, 11], in different settings.

Given an aggregation function $F = \left(F^{(k)}\right)_{k \in \mathbb{N}}$ and a profile $V = (v_i^a)$, our proposal of clustering with respect to a subset of alternatives $\emptyset \neq Y \subseteq X$ consists of a sequential process addressed by the following stages:

1. The initial clustering is $\mathcal{A}_0^Y = \{\{1\}, \ldots, \{m\}\}$.
2. Calculate the similarities between all the pairs of agents, $S_F^Y(\{a\},\{b\})$ for all $a, b \in A$.
3. Select the two agents $a, b \in A$ that maximize $S_F^Y$ and construct the first cluster $A_1^Y = \{a, b\}$.
4. The new clustering is $\mathcal{A}_1^Y = \left(\mathcal{A}_0^Y \setminus \{\{a\},\{b\}\}\right) \cup \{A_1^Y\}$.
5. Calculate the similarities $S_F^Y(A_1^Y, \{c\})$ and take into account the previously computed similarities $S_Y(\{c\},\{d\})$, for all $\{c\},\{d\} \in \mathcal{A}_1^Y$.
6. Select the two elements of $\mathcal{A}_1^Y$ that maximize $S_F^Y$ and construct the second cluster $A_2^i$.
7. Proceed as in previous items until obtaining the next clustering $\mathcal{A}_2^i$.

The process continues in the same way until obtaining the last cluster, $\mathcal{A}_{m-1}^Y = \{A\}$.

In the case of several pairs of agents or clusters are in a tie, then proceed in a lexicographic manner in $1, \ldots, m$.

## 5 Concluding Remarks

In general, clusters are usually merged by minimizing a distance between clusters. The complete, single, average and median linkage clustering take into account the maximum, minimum, mean and median distance between elements of each cluster, respectively. In turn, centroid linkage clustering is based on the distances between the clusters centroids. In these conventional linkage clustering criteria there is a loss of information. In our proposal, clusters are merged when maximizing the consensus and, consequently, all the information is used for merging clusters.

It is important emphasizing the flexibility of our proposal. Different aggregation functions can be used for measuring the consensus in each subset of agents regarding a subset of alternatives. The good properties of the corresponding consensus measure ensure a suitable clustering procedure. Nevertheless, as further research we plan to make some comparative analysis of our proposal with other clustering procedures, and also a quality measuring of our approach, in the sense of Ackerman and Ben-David [1].

**Acknowledgments** The authors gratefully acknowledge the funding support of the Spanish *Ministerio de Economía y Competitividad* (project ECO2012-32178) and *Consejería de Educación de la Junta de Castilla y León* (project VA066U13).

## References

1. Ackerman M, Ben-David S (2009) Measures of clustering quality: a working set of axioms for clustering. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in neural information processing systems, vol 21. Curran Associates, Inc., pp 121–128
2. Alcalde-Unzu J, Vorsatz M (2013) Measuring the cohesiveness of preferences: an axiomatic analysis. Soc Choice Welfare 41:965–988
3. Alcantud JCR, de Andrés R, Cascón JM (2013) On measures of cohesiveness under dichotomous opinions: some characterizations of approval consensus measures. Inf Sci 240:45–55
4. Beliakov G, Bustince Sola H, Calvo Sánchez T (2016) A practical guide to averaging functions. Springer, Heidelberg
5. Beliakov G, Pradera A, Calvo T (2007) Aggregation functions: a guide for practitioners. Springer, Heidelberg
6. Bosch R (2005) Characterizations of voting rules and consensus measures. PhD Dissertation, Tilburg University
7. Erdamar B, García-Lapresta JL, Pérez-Román D, Sanver MR (2014) Measuring consensus in a preference-approval context. Inf Fusion 17:14–21
8. Everitt BS, Landau S, Leese M (2001) Cluster analysis, 4th edn. Oxford University Press, New York
9. García-Lapresta JL, Pérez-Román D (2011) Measuring consensus in weak orders. In: Herrera-Viedma E, García-Lapresta JL, Kacprzyk J, Nurmi H, Fedrizzi M, Zadròzny S (eds) Consensual Processes, STUDFUZZ, vol 267. Springer-Verlag, Berlin, pp 213–234
10. García-Lapresta JL, Pérez-Román D (2015) Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering. Appl Soft Comput 35:864–872
11. García-Lapresta JL, Pérez-Román D (2016) Consensus-based clustering under hesitant qualitative assessments. Fuzzy Sets Syst 292:261–273
12. García-Lapresta JL, Pérez-Román D, Falcó E (2014) Consensus reaching processes under hesitant linguistic assessments. In: Angelov P et al (eds) Intelligent systems' 2014. Advances in intelligent systems and computing, vol 322, pp 257–268
13. Gini C (1912) Variabilità e Mutabilità. Tipografia di Paolo Cuppini, Bologna
14. Grabisch M, Marichal JL, Mesiar R, Pap E (2009) Aggregation functions. Cambridge University Press, Cambridge
15. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

16. Martínez-Panero M (2011) Consensus perspectives: Glimpses into theoretical advances and applications. In: Herrera-Viedma E, García-Lapresta JL, Kacprzyk J, Nurmi H, Fedrizzi M, Zadrȯzny S (eds) Consensual Processes, STUDFUZZ, vol 267. Springer-Verlag, Berlin, pp 179–193
17. Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

# Spatial Outlier Detection Using GAMs and Geographical Information Systems

**Alfonso García-Pérez and Yolanda Cabrero-Ortega**

**Abstract** A spatial (local) outlier is a value that differs from its neighbors. The usual way in which these are detected is a complicated task, especially if the data refer to many locations. In this paper we propose a different approach to this problem that consists in considering outlying slopes in an interpolation map of the observations, as indicators of local outliers. To do this, we transfer geographical properties and tools to this task using a Geographical Information System (GIS) analysis. To start, we use two completely different techniques in the detection of possible spatial outliers: First, using the observations as *heights* in a map and, secondly, using the residuals of a robust Generalized Additive Model (GAM) fit. With this process we obtain areas of possible spatial outliers (called hotspots) reducing the set of all locations to a small and manageable set of points. Then we compute the probability of such a big slope at each of the hotspots after fitting a classical GAM to the observations. Observations with a very low probability of such slope will finally be labelled as spatial outliers.

## 1 Introduction. Spatial Outliers

A local or spatial outlier [3] or [6] is an observation that differs from its neighbors, i.e., $z(s_0)$, the value of the variable of interest $Z$ at location $s_0$, is a local outlier if it differs from $z(s_0 + \Delta s_0)$ where $\Delta s_0$ defines a neighborhood of location $s_0$.

The usual method used to detect local outliers is somewhat complicated because, first, we have to define what is a neighborhood, i.e., what is "close"; then, we have to select some locations inside the neighborhood, to compute and compare the value of $Z$ at these locations.

A. García-Pérez (✉)
Departamento de Estadística, I.O. y C.N., Universidad Nacional
de Educación a Distancia (UNED), Paseo Senda del Rey 9, 28040 Madrid, Spain
e-mail: agar-per@ccia.uned.es

Y. Cabrero-Ortega
C.A. UNED-Madrid, Madrid, Spain
e-mail: ycabrero@madrid.uned.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_31

In the first part of the paper we propose two novel techniques based on a GIS for easily and quickly detect possible local outliers. The first one, developed in Sect. 2, is based on making a geographical map where the *heights* of the ground correspond to the observations. This map of separate heights is completed by means of a Triangulated Irregular Network (TIN) interpolation. Once the geographical map has been made, local outliers are easily identified as hills with big slopes.

The second technique, developed in Sect. 3, consists in fitting a robust GAM to the observations. Then, we do the previous process (interpolation plus detection of outlying slopes) with the residuals of this robust fit.

These ideas have been previously used (with some variants) in [5, 10, 12]. Here we extend their ideas considering a more general model, a GAM one, because this is the model usually considered in a fit of spatial data.

Once identified possible local outliers, we compute, in Sect. 4, the probability of such an extreme slope according to a model fitted to the data. If, according to this model (i.e., assuming that the model is correct), the probability of such extreme slope is small, the hotspot is labelled as a local outlier.

## 2 Spatial Outlier Detection by Interpolation

We propose, first, to interpolate the observations $z(s_i)$ using a TIN interpolation, that is implemented in Quantum GIS (QGIS), and that essentially means to interpolate the observations with triangles. Then we use the Geographic Resources Analysis Support System (GRASS) to compute the slopes of all the triangles obtained with the previous TIN interpolation. Finally, we reclassify the slopes, using GRASS grass again, looking for outlying slopes. All locations with big slopes will be considered as hotspots, i.e., potential outliers.

Other interpolation procedures could be used, such as Inverse Distance Weighting (IDW), but TIN works well for data with some relationship to other ones across the grid, that should be the kind of data usually considered in a spatial data problem, [8].

### *2.1 Multivariate Spatial Outliers*

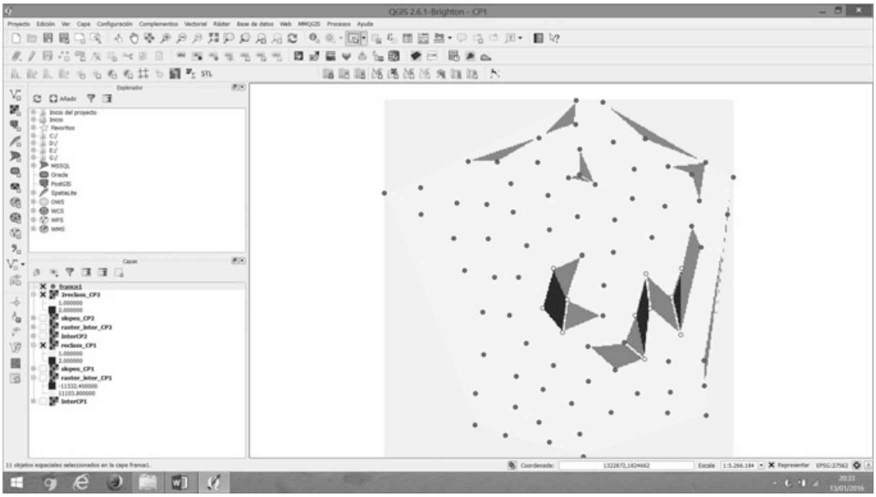
If we have multivariate observations, we first transform them into the scores obtained from a Principal Component Analysis $PC_1$, ..., $PC_p$. With this process, similar to Principal Components Regression Analysis, we can apply the previous QGIS method to each one dimensional independent variable, $PC_i$, obtaining so $p$ layers of hotspots (one layer for each $PC_i$). The intersection of all of them will be the set of possible multivariate outliers. Moreover, in this way we also have a marginal analysis for each univariate variable.

*Example 1* Let us consider Guerry data, [9], available in the R package with the same name. This data set has been analyzed in [6] and, as there, here we only use 85 departments, excluding Corsica. The two variables considered are also "population per crime against persons" (PER) and "population per crime against property" (PROP).

As we mentioned before, the descriptive process of detection of possible outliers, i.e., hotspots, consists in using QGIS, (a) incorporating first into QGIS the vectorial data, *france1.txt*, of the scores, after transforming the original observations with the two Principal Components $PC_1$ and $PC_2$; (b) computing a TIN interpolation for each new variable $PC_1$ and $PC_2$; (c) computing with GRASS the slopes from a Digital Elevation Model (DEM); (d) using again GRASS to reclassify slopes in two groups: small slopes and big slopes.

The details of the computations of all the examples in the paper are at http://www.uned.es/pfacs-estadistica-aplicada/smps.htm.

In these computations, we obtain for $PC_1$ a plot (and a table) of departments with slopes higher than 30 % and, for $PC_2$, slopes higher than 19 %. The intersection of both layers is showed in Fig. 1 where the outlying slopes (the unfilled circles) correspond to the departments Ain, Ardeche, Correze, Creuse, Indre, Isere, Jura, Loire, Rhone, Saone-et-Loire and Haute-Vienne.



11 objetos espaciales seleccionados en la capa france1.

Coordenada: 1322872,1824662 Escala 1:5.266.184

Fig. 1 Slopes reclassification ($PC_1$ and $PC_2$)

## 3 Spatial Outlier Detection by a Robust GAM

The method proposed in the previous section is an exploratory technique based only on a GIS. In this section we propose to fit a robust GAM to the spatial observations $z_i = Z(s_i)$. In this way, local large residuals will give us possible spatial outliers. We consider a GAM because this type of models is generally used for modeling spatial data.

With a GAM, [11], we assume that (univariate) observations are explained as

$$z_i = h(s_i) + h(u_{1i}) + \cdots + h(u_{ki}) + e_i \tag{1}$$

where $s_i = (x_i, y_i)$ are the coordinates of $z_i$; $u = (u_1, \ldots, u_k)$ is a vector of covariates, and $h$ is a smooth function that is expressed in terms of a basis $\{b_1, \ldots b_q\}$ as

$$h(u) = \sum_{j=1}^{q} b_j(u)\beta_j \tag{2}$$

for some unknown parameters $\beta_j$ ([15], pp. 122). The errors $e_i$ must be, as usual, i.i.d. $N(0, \sigma)$ random variables.

A key point in our proposal is to consider the coordinates $s_i = (x_i, y_i)$ of the observations $z_i$ as a covariate in model (1).

The function $h$ could be different for each covariate and, in some cases, the coordinates covariate is split into two covariates being the model

$$z_i = h_1(x_i) + h_2(y_i) + h_3(u_{1i}) + \cdots + h_{k+2}(u_{ki}) + e_i.$$

We can summarize model (1) as $z_i = H(s_i, u_{1i}, \ldots, u_{ki}) + e_i$ . This approach extends the ideas of [7] because they consider (pp. 52) a linear regression model. Also, some aspects of the papers [12] or [5] are extended in this way.

The robust GAM that we shall fit is the model proposed in [13, 14] although other possible robust GAMs could be the proposed in [1] or [4].

The robust $M$-type estimators $\hat{\boldsymbol{\beta}}$ for the GAM proposed by Wong are the solution of the following system of estimating equations

$$\sum_{i=1}^{n} \left[ w(\mu_i)\, \nu(z_i, \mu_i)\, \boldsymbol{\mu}_i' - a(\boldsymbol{\beta}) - \frac{1}{n}\mathbf{S}\boldsymbol{\beta} \right] = \mathbf{0}$$

where
$\mu_i = E[z_i|\mathbf{u}_i]$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^t$; $\boldsymbol{\mu}_i' = \partial\mu_i/\partial\boldsymbol{\beta}$; $\nu(z_i, \mu_i) = (z_i - \mu_i)/V(\mu_i)$;

$$w(\mu_i) = \frac{1}{E[\varphi_c'((z_i - \mu_i)/V^{1/2}(\mu_i))]}$$

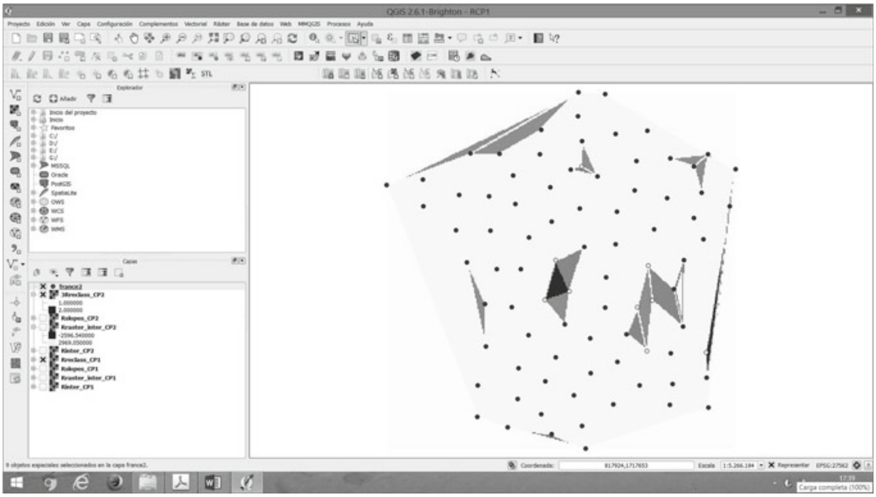
$$a(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} E_{z_i|\mathbf{u}_i}[\nu(z_i, \mu_i)]\, w(\mu_i)\, \boldsymbol{\mu}_i'$$

$\varphi_c$ the Huber-type function with tuning constant $c$, and $\mathbf{S} = 2\lambda\mathbf{D}$, being $\lambda$ a smoothing parameter and $\mathbf{D}$ a pre-specified penalty matrix.

The previous system of estimating equations, hence, is formed by the robust quasi-likelihood equations introduced in [2], plus the usual penalized GAM part.

After we have a good fit, the residuals of this fit, i.e., the differences between the observed and the predicted values, will help us to detect possible spatial outliers. To do this we compute the residuals (or the scores of the residuals if $\mathbf{z}_i(s_0)$ is multivariate), we incorporate them into QGIS and we follow the same process than in the previous section: A TIN interpolation, the slopes obtained with GRASS and, finally, a reclassification with GRASS looking for outlying slopes.

*Example 2* Let us consider Guerry data again, [9]. We first fit a robust GAM [13, 14] for each dependent variable, PER and PROP, and we compute the residuals for each fit. We then compute the scores of these residuals and, again with QGIS, we obtain departments with slopes both, higher than 30 % for $PC_1$ and higher than 13 % for $PC_2$, Fig. 2. The hotspots obtained correspond to the departments Hautes-Alpes, Ardeche, Creuse, Indre, Loire, Rhone, Saone-et-Loire, Seine and Haute-Vienne.



9 objetos espaciales seleccionados en la capa france2.

**Fig. 2** Slopes reclassification of the scores of the residuals ($PC_1$ and $PC_2$)

# 4 Identification of Spatial Outliers

With the procedures considered in the two previous sections we obtain a set of possible local outliers. In this section we compute, mathematically, if the behavior around a hotspot is very unlikely or not to label it as an actual spatial outlier, computing the probability of obtaining an slope as big as the one obtained at a given location $s_0$. Considering the framework of the last section, a large (positive or negative) slope, i.e., a large *derivative* of function $H$ ($h$ in fact) at $s_0$ will give us a good idea if $z(s_0)$ is a local outlier or not.

To compute the probabilities of large slopes at the hotspots previously identified, we first fit a classical GAM. We consider now a classical GAM fit instead of a robust one to magnify theirs slopes because the classical model will be more sensitive than the robust and the slopes less soft. Also because we know the (asymptotic) distribution of the estimators of the parameters in a classical GAM but not in the robust one.

From a mathematical point of view, the slope at a point $s_0$ in the direction $v$ is stated as the directional derivative along $v$ (unit vector) at $s_0$.

If we represent, as usual, by $D_v h(s_0)$ the collection of directional derivatives of function $h$ (assuming that it is differentiable) along all directions $v$ (unit vectors) at $s_0$ and by $MS$ the *maximum slope*, i.e., $MS(s_0) = \sup_v |D_v h(s_0)|$, we compute the probability of obtaining the observed maximum slope $ms(s_0)$, i.e., $P\{MS(s_0) \geq ms(s_0)\}$. If this probability is low (for instance lower than 0.05), we shall label $z(s_0)$ as a local outlier (more formally, we could say that we are rejecting the hypothesis of being zero the slope at $s_0$, i.e., that $z(s_0)$ is not a local outlier) and, as the smaller the probability, the greater should be considered $z(s_0)$ as a local outlier.

Because we have assumed that the smooth function $h$ has a representation in terms of a basis, (2), the slope will depend on the estimators of the parameters $\beta_j$, estimators that are approximately normal distributed ([15], pp. 189) if the $z_i$ are normal.

From vector calculus, we known that the largest value for the slope at a location $s_0$ is gradient norm, i.e.,

$$MS(s_0) = \sup_v |D_v h(s_0)| = ||\nabla h(s_0)|| = \sqrt{\left( \frac{\partial}{\partial x} h(x, y) \Big|_{s_0} \right)^2 + \left( \frac{\partial}{\partial y} h(x, y) \Big|_{s_0} \right)^2}$$

and because $h$ is expressed in term of a basis, the probability that we have to compute refers to the random variable

$$\sqrt{\left( \sum_{j=1}^{q} \frac{\partial}{\partial x} b_j(s_0) \cdot \widehat{\beta_j} \right)^2 + \left( \sum_{j=1}^{q} \frac{\partial}{\partial y} b_j(s_0) \cdot \widehat{\beta_j} \right)^2} \tag{3}$$

If this is low, $z(s_0)$ will be labelled as a local outlier.

## 4.1 Cubic Regression Splines

We shall use a cubic regression splines to explain function $h$ in the fit of a GAM to the observations $z_i$. For this aim we shall use the R function `gam` of the R package `mgcv`. The cubic spline function, with $k$ knots $v_1, \ldots, v_k$, that we fit ([15], pp. 149–150) is ($v_j \leq v \leq v_{j+1}$)

$$P(v) = \frac{v_{j+1} - v}{h_j}\,\beta_j + \frac{v - v_j}{h_j}\,\beta_{j+1} + \left[\frac{(v_{j+1} - v)^3}{h_j} - h_j\,(v_{j+1} - v)\right]\frac{\delta_j}{6}$$

$$+ \left[\frac{(v - v_j)^3}{h_j} - h_j\,(v - v_j)\right]\frac{\delta_{j+1}}{6}$$

where $h_j = v_{j+1} - v_j$, $j = 1, \ldots, k-1$ and $\delta_j = P''(v_j)$.

The first derivative of $P$ (partial derivative in formula (3)) is

$$P'(v) = \frac{\beta_{j+1} - \beta_j}{h_j} + \left[-\frac{3(v_{j+1} - v)^2}{h_j} + h_j\right]\frac{\delta_j}{6} + \left[\frac{3(v - v_j)^2}{h_j} - h_j\right]\frac{\delta_{j+1}}{6}$$

and considering as knots the locations, $v_j$,

$$P'(v_j) = \frac{\beta_{j+1} - \beta_j}{h_j} - \frac{\delta_j\,h_j}{3}.$$

If the term $\delta_j\,h_j/3$ is negligible, we have to compute the probabilities,

$$P\left\{(\hat{\beta}_{j+1} - \hat{\beta}_j)/h_j > \text{ observed slope}\right\}$$

based on a normal model because ([15], pp. 189) $\hat{\beta}_j$ is approximately normal distributed with mean $\beta_j$.

**Table 1** Probability of a big slope for both variables

| | | Probability | |
|---|---|---|---|
| Dept | Department | PER | PROP |
| 5 | Hautes-Alpes | 0.08677979 | 0.734663 |
| 1 | Ain | 0.7796545 | 0.9039119 |
| 7 | Ardeche | 0.08590459 | 0.5845837 |
| 19 | Correze | 0.8543756 | 0.968079 |
| 23 | Creuse | 0.3344432 | 0.8536806 |
| 36 | Indre | 0.8043197 | 0.9364876 |
| 38 | Isere | 0.2926037 | 0.7874324 |
| 39 | Jura | 0 | 0.0062001 |
| 42 | Loire | 0.5497284 | 0.8805521 |
| 69 | Rhone | 0 | 0.365532 |
| 71 | Saone-et-Loire | 0.45913 | 0.8109866 |
| 75 | Seine | 0 | 0 |
| 87 | Haute-Vienne | 0.01982465 | 0.6981038 |

*Example 3* Let us consider Guerry data again. The set of all departments detected as possible outliers for, at least, one of the two methods explained in Sects. 2 and 3, together with the probabilities of such slopes (i.e., the p-values of the bilateral test of the null hypothesis $H_0 : \beta_{j+1} - \beta_j = 0$), are in Table 1.

Hence, we can label as spatial outliers the observations at Jura, Rhone and Seine. As is remarked in [6], Seine (together with Ain, Haute-Loire and Creuse) is a global outlier and a local one.

Hence, if we do not consider the Department of Seine (because is a global outlier) we have two departments that can be considered as spatial outliers: Jura and Rhone, two departments in what is called the Rhône-Alpes area, i.e., the same result than in [6].

**Acknowledgment** This work is partially supported by Grant MTM2012-33740.

## References

1. Alimadad A, Salibian-Barrera M (2011) An outlier-robust fit for generalized additive models with applications to disease outbreak detection. J Am Stat Assoc 106:719–731
2. Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. J Am Stat Assoc 96:1022–1030
3. Cressie NAC (1993) Statistics for spatial data. Wiley, New York
4. Croux C, Gijbels I, Prosdocimi I (2012) Robust estimation of mean and dispersion functions in extended generalized additive models. Biometrics 68:31–44
5. Felicísmo AM (1994) Parametric statistical method for error detection in digital elevation models. ISPRS J Photogramm Remote Sens 49:29–33
6. Filzmoser P, Ruiz-Gazen A, Thomas-Agnan C (2014) Identification of local multivariate outliers. Stat Papers 55:29–47
7. Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression. The analysis of spatially varying relationships. Wiley, New York
8. Franke R (1982) Scattered data interpolation: tests of some methods. Math Comput 38:181–200
9. Guerry A-M (1833) Essai sur la statistique morale de la France. Crochard, Paris. English translation: HP Whitt and VW Reinking, Edwin Mellen Press, Lewiston, 2002
10. Hannah MJ (1981) Error detection and correction in digital terrain models. Photogramm Eng Remote Sens 47:63–69
11. Hastie T, Tibshirani R (1990) Generalized additive models. Chapman & Hall, London
12. Liu H, Jezek KC, OKelly ME (2001) Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. Int J Geogr Inf Sci 15:721–741
13. Wong KW (2010) Robust Estimation for Generalized Additive Models. MA Thesis, Department of Statistics, The Chinese University of Hong Kong
14. Wong RKW, Yao F, Lee TCM (2014) Robust estimation for generalized additive models. J Comput Graph Stat 23:270–289
15. Wood SN (2006) Generalized additive models: an introduction with R. Chapman & Hall/CRC Press

# Centering and Compound Conditionals Under Coherence

**Angelo Gilio, David E. Over, Niki Pfeifer and Giuseppe Sanfilippo**

**Abstract** There is wide support in logic, philosophy, and psychology for the hypothesis that the probability of the indicative conditional of natural language, $P(\textit{if } A \textit{ then } B)$, is the conditional probability of $B$ given $A$, $P(B|A)$. We identify a conditional which is such that $P(\textit{if } A \textit{ then } B) = P(B|A)$ with de Finetti's conditional event, $B|A$. An objection to making this identification in the past was that it appeared unclear how to form compounds and iterations of conditional events. In this paper, we illustrate how to overcome this objection with a probabilistic analysis, based on coherence, of these compounds and iterations. We interpret the compounds and iterations as conditional random quantities, which sometimes reduce to conditional events, given logical dependencies. We also show, for the first time, how to extend the inference of centering for conditional events, inferring $B|A$ from the conjunction $A$ and $B$, to compounds and iterations of both conditional events and biconditional events, $B||A$, and generalize it to $n$-conditional events.

Shared first authorship (all authors contributed equally to this work).

A. Gilio
University of Rome "La Sapienza", Rome, Italy
e-mail: angelo.gilio@sbai.uniroma1.it

D.E. Over
University of Durham, Durham, UK
e-mail: david.over@durham.ac.uk

N. Pfeifer
Ludwig-Maximilians-University Munich, Munich, Germany
e-mail: niki.pfeifer@lmu.de

G. Sanfilippo (✉)
University of Palermo, Palermo, Italy
e-mail: giuseppe.sanfilippo@unipa.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_32

# 1 Introduction

There is wide agreement in logic and philosophy that the indicative conditional of natural language, *if A then B*, cannot be adequately represented as the material conditional of binary logic, logically equivalent to $\overline{A} \vee B$ (*not-A or B*) [8]. Psychological studies have also shown that ordinary people do not judge the probability of *if A then B*, $P(\textit{if } A \textit{ then } B)$, to be the probability of the material conditional, $P(\overline{A} \vee B)$, but rather tend to assess it as the conditional probability of $B$ given $A$, $P(B|A)$, or at least to converge on this assessment [3, 10, 11, 24, 27]. These psychological results have been taken to imply [3, 9, 13, 22, 24], that *if A then B* is best represented, either as the probability conditional of Adams [2], or as the *conditional event* $B|A$ of de Finetti [5, 6], the probability of which is $P(B|A)$. We will here adopt the latter view and base our analysis on conditional events and coherence (for analyses on categorical syllogisms and the square of opposition under coherence see [19, 25]). Given two events $B$ and $A$, with $A \neq \bot$, the *conditional event* $B|A$ is defined as a three-valued logical entity which is *true* if $AB$ (i.e., $A \wedge B$) is true, *false* if $\overline{B}A$ is true, and *void* if $A$ is false.

In the above and what follows, we use the same symbols, for instance $A$, to refer to the (unconditional) event $A$ and its indicator. One possible objection to holding that $P(\textit{if } A \textit{ then } B) = P(B|A)$ is that it is supposedly unclear how this relation extends to compounds of conditionals and makes sense of them [7, 8, 28]. Yet consider:

$$\overbrace{\textit{She is angry}}^{a} \text{ if } \overbrace{\textit{her son gets a } \mathrm{B}}^{b} \text{ and } \overbrace{\textit{she is furious}}^{f} \text{ if } \overbrace{\textit{he gets a } \mathrm{C}}^{c}. \qquad (1)$$

The above conjunction appears to make sense, as does the following seemingly even more complex conditional construction [7]:

If *she is angry* if *her son gets a* B, then *she is furious* if *he gets a* C. (2)

We will show below, in reply to the objection, how to give sense to (1) and (2) in terms of *compound conditionals*. Specifically, we will interpret (1) as a *conjunction* of two conditionals ($a|b$ and $f|c$) and (2) in terms of a *conditional* whose antecedent ($a|b$) and consequent ($f|c$) are both conditionals (if $a|b$, then $f|c$). But we note first that (2) validly follows from (1) by the form of inference we will call *centering* (which is often termed "conjunctive sufficiency") [21], when this is extended to compounds of conditionals (see Sect. 2). We point out that our framework is quantitative rather than a logical one. Indeed in our approach, syntactically conjoined and iterated conditionals in natural language are analyzed as conditional random quantities, which can sometimes reduce to conditional events, given logical dependencies [15, 18]. For instance, the biconditional event $A||B$, which we will define by $(B|A) \wedge (A|B)$, reduces to the conditional event $(A \wedge B)|(A \vee B)$. Moreover, the notion of *biconditional centering* will be given. We will also introduce the notion of $n$-conditional centering (see Sect. 3). Finally, in Sect. 4 we will give some remarks on future work which will involve counterfactuals.

## 2 Centering

There is *one-premise centering*: inferring *if $A$ then $B$* from the single premise $A \wedge B$. And *two-premise centering*: inferring *if $A$ then $B$* from the two separate premises $A$ and $B$. Centering is valid for quite a wide range of conditionals [23]. It is clearly valid for the material conditional, since *not-$A$ or $B$* must be true when *$A$ and $B$* is true. It is also valid for Lewis conditional *if $A$ then $B$* [21], which holds, roughly, when $B$ is true in the closest world in which $A$ is true. In [21] Lewis has a semantic condition of centering, which states that the actual world is the closest world to itself. The characteristic inference rule for this semantic condition is what we are also calling centering. This rule is probabilistically valid, *p-valid*, for the conditional event, i.e. the premise *p-entails* the conclusion.

A (p-consistent) set of premises p-entails a conclusion if and only if the conclusion must have probability one when all the premises have probability one [16]. Clearly, one-premise centering is p-valid, indeed the p-entailment of $B|A$ from $A \wedge B$ follows by observing that $P(A \wedge B) = P(A)P(B|A)$ and so $P(A \wedge B) \leq P(B|A)$: if $P(A \wedge B) = 1$, then $P(B|A) = 1$. Two-premise centering is also clearly p-valid, as it is p-valid to infer $A \wedge B$ from $A$ and $B$, and then one-premise centering can be used to infer $B|A$: if $P(A) = x$ and $P(B) = y$, coherence requires that $P(A \wedge B)$ has to be in the interval $[\max\{0, x + y - 1\}, \min\{x, y\}]$, with $P(A \wedge B) \leq P(B|A)$. Therefore, the set of premises $\{A, B\}$ p-entails $B|A$: if $P(A) = P(B) = 1$, it follows $P(A \wedge B) = P(B|A) = 1$. We will give here a probabilistic analysis of two-premise centering when the premises are conditionals, $A|H$, $B|K$, and the conclusion is an iterated conditional, $(B|K)|(A|H)$. We recall that in the approach given in [14, 15, 18] any conditional event $A|H$ can be seen as the random quantity $AH + xH^c \in \{1, 0, x\}$, where $x = P(A|H)$. In the same papers the notions of conjunction and iterated conditioning for two conditional events are studied. We give the notion of conjunction below.

**Definition 1** (*Conjunction*) Given any pair of conditional events $A|H$ and $B|K$, with $P(A|H) = x$, $P(B|K) = y$, we define their conjunction as the conditional random quantity

$$(A|H) \wedge (B|K) = min\,\{A|H, B|K\}\,|\,(H \vee K) = (A|H) \cdot (B|K)\,|\,(H \vee K)\,.$$

Based on the betting scheme the compound conditional $(A|H) \wedge (B|K)$ coincides with $1 \cdot AHBK + x \cdot \overline{H}BK + y \cdot AH\overline{K} + z \cdot \overline{H}\,\overline{K}$, where $z$ is the *prevision* of the random quantity $[(A|H) \wedge (B|K)]$, denoted by $\mathbb{P}[(A|H) \wedge (B|K)]$. Notice that $z$ represents the amount you agree to pay, with the proviso that you will receive the quantity $(A|H) \wedge (B|K)$. By linearity of prevision, if $P(H \vee K) > 0$ it holds that [20]

$$\mathbb{P}[(A|H) \wedge (B|K)] = \frac{P(AHBK) + P(A|H)P(\overline{H}BK) + P(B|K)P(AH\overline{K})}{P(H \vee K)}\,.$$

For examples see [15] and [20]. We remark that in the setting of coherence de Finetti's notion of prevision $\mathbb{P}$, which corresponds to the notion of expected value, may be evaluated in a direct way. We recall the following result [18]:

**Theorem 1** *Given any coherent assessment* $(x, y)$ *on* $\{A|H, B|K\}$*, with* $A, H, B, K$ *logically independent, and with* $H \neq \emptyset, K \neq \emptyset$*, the extension* $z = \mathbb{P}[(A|H) \wedge (B|K)]$ *is coherent if and only if the Fréchet-Hoeffding bounds are satisfied:*

$$max\{x + y - 1, 0\} = z' \ \leq \ z \ \leq \ z'' = min\{x, y\}\,. \tag{3}$$

*Remark 1* From (3) it holds that $0 \leq z' \leq z'' \leq 1$ for every coherent assessment $(x, y)$. Moreover, if $x = 1, y = 1$, then $z' = z'' = 1$. Thus, $z = 1$ is the unique coherent extension. Then, by adopting the usual language, we say that

$$\{A|H, B|K\} \models_p (A|H) \wedge (B|K)\,, \tag{4}$$

where "$\models_p$" denotes p-entailment. We call this inference rule "*And* rule for conditional events."

Now, we recall the notion of iterated conditioning.

**Definition 2** (*Iterated conditioning*) Given any pair of conditional events $A|H$ and $B|K$, the iterated conditional $(B|K)|(A|H)$ is the conditional random quantity $(B|K)|(A|H) = (B|K) \wedge (A|H) + \mu \bar{A}|H$, where $\mu = \mathbb{P}[(B|K)|(A|H)]$.

Notice that, in the context of betting scheme, $\mu$ represents the amount you agree to pay, with the proviso that you will receive the quantity

$$(B|K)|(A|H) = \begin{cases} 1, & \text{if } AHBK \text{ true,} \\ 0, & \text{if } AH\bar{B}K \text{ true,} \\ y, & \text{if } AH\bar{K} \text{ true,} \\ \mu, & \text{if } \bar{A}H \text{ true,} \\ x + \mu(1 - x), & \text{if } \bar{H}BK \text{ true,} \\ \mu(1 - x), & \text{if } \bar{H}\bar{B}K \text{ true,} \\ z + \mu(1 - x), & \text{if } \bar{H}\bar{K} \text{ true.} \end{cases} \tag{5}$$

We recall the following product formula [15]

**Theorem 2** (Product formula) *Given any assessment* $x = P(A|H)$, $\mu = \mathbb{P}[(B|K)|(A|H)]$, $z = \mathbb{P}[(B|K) \wedge (A|H)]$, *if* $(x, y, z)$ *is coherent, then* $z = \mu \cdot x$.

The result in Theorem 2 can be obtained by applying the linearity of prevision [18]; indeed by linearity:

$$\mathbb{P}[(B|K)|(A|H)] = \mu = \mathbb{P}[(B|K) \wedge (A|H)] + \mu P(\bar{A}|H) = z + \mu(1 - x)\,, \tag{6}$$

from which it follows $z = \mu \cdot x$, that is

$$\mathbb{P}[(B|K) \wedge (A|H)] = \mathbb{P}[(B|K)|(A|H)]P(A|H)\,. \tag{7}$$

Moreover, by taking into account (6), $(B|K)|(A|H)$ coincides with

$$1\,AHBK + yAH\overline{K} + (x + \mu(1-x))\,\overline{H}BK + \mu(1-x)\,\overline{H}\,\overline{B}K + \mu\,(\overline{A}H \vee \overline{H}\,\overline{K}).$$

*One-premise* centering is p-valid, indeed the p-entailment of $(B|K)|(A|H)$ from $(B|K) \wedge (A|H)$ follows from (7) by observing that

$$\mathbb{P}[(B|K) \wedge (A|H)] \leq \mathbb{P}[(B|K)|(A|H)], \tag{8}$$

therefore: if $\mathbb{P}[(B|K) \wedge (A|H)] = 1$, then $\mathbb{P}[(B|K)|(A|H)] = 1$.
*Two-premise* centering is also p-valid; indeed, it is p-valid to infer $(B|K) \wedge (A|H)$ from $A|H$ and $B|H$, and then one-premise centering can be used to infer $(B|K)|(A|H)$: by applying Theorem 1 with $x = P(A|H) = 1$ and $y = P(B|K) = 1$, it follows that the extension $z = \mathbb{P}[(B|K) \wedge (A|H)]$ is coherent if and only if $z = 1$. Therefore, based on (8), the set of (p-consistent) premises $\{A|H, B|K\}$ p-entails $(B|K)|(A|H)$: if $P(A|H) = P(B|K) = 1$, then $\mathbb{P}[(A|H) \wedge (B|K)] = \mathbb{P}[(B|K)|(A|H)] = 1$.

*Remark 2* If we only assign the values $x = P(A|H)$ and $y = P(B|K)$, by Theorem 1, we obtain $z \in [\max\{0, x + y - 1\}, \leq \min\{x, y\}]$, where $z = \mathbb{P}[(A|H) \wedge (B|K)]$. Then, by assuming $x > 0$, by Theorem 2 it follows that the extension $\mu = \mathbb{P}[(B|K)|(A|H)]$ is coherent if and only if $\mu \in [\mu', \mu'']$, where $\mu' = \max\left\{0, \frac{x+y-1}{x}\right\}$ and $\mu'' = \min\left\{1, \frac{y}{x}\right\}$. When $x = 0$ (by the penalty criterion) we can prove that $[\mu', \mu''] = [0, 1]$.

## 3 Biconditional and *n*-Conditional Centering

In classical logic the biconditional $A \leftrightarrow B$ (defined by $\overline{(A \vee B)} \vee (A \wedge B)$) can be represented by the conjunction of the two material conditionals $\overline{A} \vee B$ and $\overline{B} \vee A$. Therefore, $\{\overline{A} \vee B, \overline{B} \vee A\} \models A \leftrightarrow B$, which is called *biconditional introduction rule*. With the material conditional interpretation of a conditional, the biconditional $A \leftrightarrow B$ represents the conjunction of the two conditionals *if A then B* and *if B then A*. In this section, we present an analogue in terms of conditional events, by also giving a meaning to the conjunction of two conditional events $A|B$ and $B|A$.

From centering it follows that $\{A, B\} \models_p B|A$ and $\{A, B\} \models_p A|B$. Then, from $P(A) = P(B) = 1$ it follows that $P(B|A) = P(A|B) = 1$, which we denoted by: $\{A, B\} \models_p \{A|B, B|A\}$. Thus, by applying (4) with $H = B$ and $K = A$, we obtain $\{A|B, B|A\} \models_p (A|B) \wedge (B|A)$ (which we call *biconditional introduction* rule). Then, by transitivity

$$\{A, B\} \;\models_p\; (A|B) \wedge (B|A)\,. \tag{9}$$

In a similar way, we can prove that

$$A \wedge B \ \models_p \ (A|B) \wedge (B|A)\,. \tag{10}$$

We recall that the conditional event $(A \wedge B) \,|\, (A \vee B)$, denoted by $A||B$, captures the notion of the *biconditional event*, which has been seen as the conjunction of two conditionals with the same truth table as the "defective" biconditional discussed in [12]; see also [11]. We have

**Theorem 3** *Given two events* $A$ *and* $B$ *it holds that:* $(A|B) \wedge (B|A) = (A \wedge B)|(A \vee B) = A||B$.

*Proof* We note that $(A|B) \wedge (B|A) = \min(A|B, B|A)|(A \vee B) = AB + \mu \cdot \overline{A}\,\overline{B}$, where $\mu = \mathbb{P}[(A|B) \wedge (B|A)]$; we also observe that $(A \wedge B)|(A \vee B) = AB + p \cdot \overline{A}\ \overline{B}$, where $p = P[(A \wedge B)|(A \vee B)]$. Then, under the assumption that "$(A \vee B)$ is true", the two random quantities $(A|B) \wedge (B|A)$ and $(A \wedge B)|(A \vee B)$ coincide. By coherence (see [18, Theorem 4]) it follows that these two random quantities coincide also under the assumption that "$(A \vee B)$ is false", that is $\mu$ and $p$ coincide. Therefore, $(A|B) \wedge (B|A) = (A \wedge B)|(A \vee B)$. □

Therefore, based on Theorem 3, we can now really interpret the biconditional event $A||B$ as the conjunction of the two conditionals $(B|A)$ and $(A|B)$. Moreover, equations (9) and (10) represent what we call *two-premise biconditional centering* and *one-premise biconditional centering* respectively, that is $\{A, B\} \models_p A||B$ and $A \wedge B \models_p A||B$.

Though in classical logic $\{\overline{A}, \overline{B}\} \models (A \leftrightarrow B)$, the analogue does not hold in our approach, since we do not have p-entailment of $A||B$ from $\overline{A}, \overline{B}$, indeed if $P(\overline{A}) = P(\overline{B}) = 1$, then $P(A \vee B) = 0$ and therefore $P(A||B) = P((A \wedge B)|(A \vee B)) \in [0, 1]$. The biconditional event $A||B$ is of interest to psychologists because there is evidence that children go through a developmental stage in which they judge that $P$(*if A then B*) $= P[(A \wedge B)|(A \vee B)]$, with this judgment being replaced by *P(if A then B)* = $P(B|A)$ as they grow older [12]. We recall that, given two conditional events $A|H$ and $B|K$, their quasi conjunction is defined as the conditional event $Q(A|H, B|K) = [(AH \vee \overline{H}) \wedge (BK \vee \overline{K})]|(H \vee K)$. Quasi conjunction is a basic notion in the work of Adams [1] and plays a role in characterizing entailment from a conditional knowledge base (see also [4]). We recall that in [17] $A||B$ was interpreted by the quasi conjunction of $A|B$ and $B|A$, by obtaining $A||B = Q(A|B, B|A) = (A \wedge B)|(A \vee B)$. In the same paper the following probabilistic rule is given. Let $(x, y)$ be any coherent assessment on $\{A|B, B|A\}$; then, the probability assessment $z = P(A||B)$ is a coherent extension of $(x, y)$ if and only if

$$z = T_0^H(x, y) = \begin{cases} 0, & (x = 0 \vee y = 0)\,, \\ \frac{xy}{x+y-xy} = \frac{1}{\frac{1-x}{x} + \frac{1-y}{y} + 1}, & (x \neq 0 \wedge y \neq 0)\,, \end{cases} \tag{11}$$

where $T_0^H(x, y)$ is the Hamacher t-norm, with parameter $\lambda = 0$. Of course, two-premise centering for the biconditional event directly follows by instantiating (11) with $x = y = 1$. In [17] the notion of biconditional event has been generalized by defining the $n$-conditional event. Given $n$ (non-impossible) events $A_1, \dots, A_n$, the associated $n$-conditional event is given by

$$\begin{aligned} A_1||A_2||\cdots||A_n &= Q(A_2|A_1, A_3|A_2, \dots, A_n|A_{n-1}, A_1|A_n) = \\ &= (A_1 \wedge \cdots \wedge A_n) \,|\, (A_1 \vee \cdots \vee A_n)\,. \end{aligned}$$

Then, by recalling that the extension of a t-norm $T$ in $[0, 1]^n$ is defined as

$$T(p_1, p_2, \dots, p_n) = \begin{cases} T(T(p_1, \dots, p_{n-1}), p_n), & n > 2, \\ T(p_1, p_2), & n = 2, \end{cases}$$

and based on [17, Proposition 3] we obtain

**Theorem 4** *Given any coherent assessment $(p_1, p_2, \dots, p_n)$ on $\{A_1, A_2, \dots, A_n\}$, for every $k = 2, \dots, n$, the extension $z_k = P(A_1||A_2||\cdots||A_k)$ of $(p_1, p_2, \dots, p_k)$ is coherent if and only if*

$$z_k = T_0^H(p_1, p_2, \dots, p_k) = \begin{cases} 0, & p_i = 0 \textit{ for at least one } i, \\ \frac{1}{\sum_{i=1}^{k} \frac{1-p_i}{p_i} + 1}\,, & p_i > 0 \textit{ for } i = 1, \dots, k\,. \end{cases} \tag{12}$$

By formula (12), if $p_1 = p_2 = \ldots = p_n = 1$, then $z = 1$, that is $A_1, A_2, \dots, A_n \models_p A_1||A_2||\cdots||A_n$, which we call *n-premise n-conditional centering*. As a further observation we note that, by applying the And rule to the events $A_1, \dots, A_n$, we obtain the *one-premise n-conditional centering* $A_1 \wedge A_2 \wedge \ldots \wedge A_n \models_p A_1||A_2||\cdots||A_n$.

## 4 Conclusion

In this paper, we have illustrated a probabilistic analysis of the conjunction and iteration of conditional events, and of the centering inference for these conjunctions and iterations. We see this analysis as relevant to the conjunction and iteration of indicative conditionals in natural language. It is often argued that there are deep differences between indicative and counterfactual conditionals in natural language. For example, the indicative conditional *If Oswald did not kill Kennedy, someone else did* seems very different from the counterfactual conditional *If Oswald had not killed Kennedy then someone else would have* [8]. However, we will consider extending our approach in future work to counterfactuals (see [20] for points relevant to this and see [26] for an experimental study comparing systematically counterfactuals and indicative conditionals under coherence).

**Acknowledgments** We thank the anonymous referees for useful comments. We thank Villa Vigoni (*Human Rationality: Probabilistic Points of View*). N. Pfeifer is supported by his DFG project PF 740/2-2 (within the SPP1516). G. Sanfilippo is supported by the INdAM–GNAMPA Projects (2016 Grant U 2016/000391 and 2015 grant U 2015/000418).

## References

1. Adams EW (1975) The logic of conditionals. Reidel, Dordrecht
2. Adams EW (1998) A primer of probability logic. CSLI, Stanford
3. Baratgin J, Over DE, Politzer G (2013) Uncertainty and the de Finetti tables. Thinking Reasoning 19:308–328
4. Benferhat S, Dubois D, Prade H (1997) Nonmonotonic reasoning, conditional objects and possibility theory. Artif Intell 92:259–276
5. de Finetti B (1936/1995) The logic of probability. Philos Stud 77:181–190
6. de Finetti B (1937/1980) Foresight: its logical laws, its subjective sources. In: Studies in subjective probability. Krieger, Huntington, pp 55–118
7. Douven I (2015) On de Finetti on iterated conditionals. Technical report, CNRS
8. Edginton D (2014) Indicative conditionals. Stanford Encyclopedia of Philosophy
9. Evans JSBT, Over DE (2004) If. Oxford University Press, Oxford
10. Evans JSBT, Handley SJ, Over DE (2003) Conditionals and conditional probability. JEP:LMC 29(2):321–355
11. Fugard AJB, Pfeifer N, Mayerhofer B, Kleiter GD (2011) How people interpret conditionals: shifts towards the conditional event. JEP:LMC 37(3):635–648
12. Gauffroy C, Barrouillet P (2009) Heuristic and analytic processes in mental models for conditionals. Dev Rev 29:249–282
13. Gilio A, Over DE (2012) The psychology of inferring conditionals from disjunctions: a probabilistic study. J Math Psychol 56:118–131
14. Gilio A, Sanfilippo G (2013) Conditional random quantities and iterated conditioning in the setting of coherence. In: van der Gaag LC (ed) ECSQARU 2013, vol 7958, LNCS. Springer, Berlin, Heidelberg, pp 218–229
15. Gilio A, Sanfilippo G (2013) Conjunction, disjunction and iterated conditioning of conditional events. Synergies of Soft Computing and Statistics for Intelligent Data Analysis, volume 190 of AISC. Springer, Berlin, pp 399–407
16. Gilio A, Sanfilippo G (2013) Probabilistic entailment in the setting of coherence: the role of quasi conjunction and inclusion relation. Int J Approx Reason 54(4):513–525
17. Gilio A, Sanfilippo G (2013) Quasi conjunction, quasi disjunction, t-norms and t-conorms: probabilistic aspects. Inf Sci 245:146–167
18. Gilio A, Sanfilippo G (2014) Conditional random quantities and compounds of conditionals. Stud Logica 102(4):709–729
19. Gilio A, Pfeifer N, Sanfilippo G (2016) Transitivity in coherence-based probability logic. J Appl Logic 14:46–64
20. Kaufmann S (2009) Conditionals right and left: probabilities for the whole family. J Philos Logic 38:1–53
21. Lewis D (1973) Counterfactuals. Blackwell, Oxford
22. Oaksford M, Chater N (2007) Bayesian rationality: the probabilistic approach to human reasoning. Oxford University Press, Oxford
23. Over DE (in press) Causation and the probability of causal conditionals. In: The Oxford handbook of causal reasoning. OUP, Oxford
24. Pfeifer N (2013) The new psychology of reasoning: a mental probability logical perspective. Thinking Reasoning 19(3–4):329–345
25. Pfeifer N, Sanfilippo G Square of opposition under coherence. In: This issue
26. Pfeifer N, Stöckle-Schobel R (2015) Uncertain conditionals and counterfactuals in (non-)causal settings. In: Proceedings of the 4th European conference on cognitive science, vol 1419. CEUR Workshop Proceedings, pp 651–656
27. Singmann H, Klauer KC, Over DE (2014) New normative standards of conditional reasoning and the dual-source model. Frontiers Psychol 5:Article 316
28. van Wijnbergen-Huitink J, Elqayam S, Over DE (2015) The probability of iterated conditionals. Cogn Sci 39(4):788–803

# Approximate Bayesian Methods for Multivariate and Conditional Copulae

**Clara Grazian and Brunero Liseo**

**Abstract** We describe a simple method for making inference on a functional of a multivariate distribution. The method is based on a copula representation of the multivariate distribution, where copula is a flexible probabilistic tool that allows the researcher to model the joint distribution of a random vector in two separate steps: the marginal distributions and a copula function which captures the dependence structure among the vector components. The method is also based on the properties of an approximate Bayesian Monte Carlo algorithm, where the proposed values of the functional of interest are weighted in terms of their empirical likelihood. This method is particularly useful when the likelihood function associated with the working model is too costly to evaluate or when the working model is only partially specified.

## 1 Introduction

Theoretical proposals are now available to model complex situations, thanks to the recent advances in computational methodologies and to the increased power of modern computers. In particular, there are new methods for multivariate analysis, however the goal of modelling complex multivariate structures and estimating them has not yet been reached in a completely satisfactory way.

Copula models have been introduced as probabilistic tools to describe a multivariate random vector via the marginal distributions and a copula function which captures the dependence structure among the vector components, thanks to the Sklar's theorem [1], which states that any $d$-dimensional absolutely continuous density can be uniquely represented as

$$f(x_1, \dots, x_d) = f_1(x_1) \dots f_d(x_d) c_{12\dots d}(F_1(x_1), \dots, F_d(x_d)). \tag{1}$$

C. Grazian · B. Liseo (✉)
MEMOTEF, Sapienza Università di Roma, Roma, Italy
e-mail: brunero.liseo@uniroma1.it

C. Grazian
e-mail: clara.grazian@uniroma1.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_33

While it is often straightforward to produce reliable estimates of the marginals, making inference on the dependence structure is more complex. Major areas of application include econometrics, hydrological engineering, biomedical science, signal processing and finance.

In a parametric frequentist approach to copula models, there are no broadly satisfactory methods for the joint estimation of marginal and copula parameters. The most popular method is the so called Inference From the Margins (IFM), where the parameters of the marginal distributions are estimated first, and then pseudo-observations are obtained by pluggin-in the estimates of the marginal parameters. Then inference on the copula parameters is performed using the pseudo-observations: this approach obviously does not account for the uncertainty on the estimation of the marginal parameters. Bayesian alternatives are not yet fully developed, although there are remarkable exceptions ([2, 3], among others).

In this work we consider the general problem of estimating some specific quantities of interest of a generic copula (such as, for example, tail dependence index or Spearman's $\rho$) by adopting an approximate Bayesian approach along the lines of [4]. In particular, we discuss the use of the an approximate Bayesian computation algorithm based on empirical likelihood weights (in the following, $BC_{EL}$), based on the empirical likelihood approximation of the marginal likelihood of the quantity of interest (see [5] for a complete and recent survey on empirical likelihood methods). This approach produces an approximation of the posterior distribution of the quantities of interest, based on an approximation of the likelihood function and on a Monte Carlo approximation of the posterior distribution via simulations. Our approach is general, in the sense that it could be adapted both to parametric and nonparametric modelling of the marginal distributions. Also, the use of empirical likelihood avoids the need of choosing a specific parametric copula model.

## 2 Approximate Bayesian Computation

The idea underlying likelihood-free methods (or approximate Bayesian computation, ABC) is to propose a candidate $\theta'$ and to generate a data set from the working model with parameter set to $\theta'$. If the observed and the simulated data are similar "in some way", then the proposed value is considered a good candidate to have generated the data and becomes part of the sample which will form the approximation to the posterior distribution. Conversely, if the observed and the simulated data are too different, the proposed $\theta'$ is discarded.

In order to implement a basic version of the ABC algorithm one needs to specify a set of summary statistics to make comparisons, a distance to quantify comparisons and a tolerance level $\epsilon > 0$.

The basic ABC may be inefficient, because it proposes values of $\theta$ from its prior distribution, therefore, ABC algorithms are often linked with other methods, for instance, with Markov Chain Monte Carlo (MCMC) or Sequential Monte Carlo (SMC) methods. In this work, we will focus on a different ABC approach, described in Algorithm 1 where $M$ simulations from the prior are generated; this method avoids the most expensive step in computational time, that is the generation of new data sets. A detailed description of the method is in [4]; it represents a re-sampling scheme where the proposed values are re-sampled with weights proportional to their empirical likelihood values.

```
for $i = 1$ to $M$ do
repeat
Generate $\theta_i$ from the prior distribution $\pi(\theta)$
Set the weight for $\theta_i$ as $\omega_i = L_{EL}(\theta_i; \mathbf{x})$.
end for
for $i = 1$ to $M$ do
Draw, with replacement, a value $\theta_i$ from the previous set of $M$ values using weights
$\omega_i, i = 1, \dots, M$.
end for
```

**Algorithm 1:** BC$_{EL}$ algorithm

## 3 The Proposed Approach

We propose to adapt the BC$_{EL}$ algorithm to a situation where the statistical model is only partially specified and the main goal is the estimation of a finite dimensional quantity of interest, i.e. a situation where the complete structure of the mutual dependence is considered a nuisance parameter and it is kept as general as possible. While the main interest of [4] was the approximation of the full posterior distribution of the parameters of the model, here we use the empirical likelihood (EL) approach to avoid a parametric definition of the model for the observed data and focus only on a particular functional of the distribution, which summarizes the correlation among the variables.

We assume that a data set is available in the form of a size $n \times d$ matrix $\boldsymbol{X}$, where $n$ is the sample size and $d$ is the number of variables, that is

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & x_{ij} & \dots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix}.$$

In the following, $X_{[\cdot,j]}$ will denote the $j$-th column (variable) and $X_{[i,\cdot]}$ the $i$-th row of $\boldsymbol{X}$, respectively. For each $j = 1, \dots, d$, we consider the available data information in $X_{[\cdot,j]}$ to produce an estimate of the marginal CDF of $X_{[\cdot,j]}$. Let $\boldsymbol{\lambda}_j = (\lambda_j^{(1)}, \lambda_j^{(2)}, \dots \lambda_j^{(B)}),\ j = 1, 2, \dots d$, be the posterior sample of size $B$ obtained for the distribution of $X_{[\cdot,j]}$. Notice that the vector $\boldsymbol{\lambda}_j$ can be either a sample from the posterior distribution of the parameters of the model we have adopted for $X_{[\cdot,j]}$ or a posterior sample of $CDF$'s in a nonparametric set-up.

Then we use a copula representation for estimating the multivariate dependence structure of the random vector $\boldsymbol{X}$,

$$H(x_1, \ldots, x_d) = C_\theta\big(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\big) \tag{2}$$

where $\theta$ is the parameter related to the copula function.

Estimating the copula $C_\theta(\cdot)$ can be managed either using some parametric model for the copula (such as Clayton, Gaussian, Skew-t, Gumbel, etc.) or using a nonparametric approach. In this paper, we take a nonparametric route (in many situations it is difficult to prefer a model instead of another) and we concentrate on some specific function of $C_\theta(\cdot)$, say $\varphi = T(\boldsymbol{F})$, for example the Spearman's measure of association $\rho$ between two components of $\boldsymbol{X}$, say $X_h$ and $X_j$, which is defined as the correlation coefficient among the transformed values $U_i = F_i(x_i),\ i = j, h$:

$$\rho = 12 \int_0^1 \int_0^1 \big(C(u_j, u_h) - u_h u_j\big) du_j du_h. \tag{3}$$

Its sampling counterpart $\rho_n$ is the correlation among ranks $\mathbf{R}$ and $\mathbf{S}$ of the data observed for the two variables of interest and it can be written as:

$$\rho_n = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{12}{n^2 - 1} R_i S_i \right) - 3\frac{n+1}{n-1}. \tag{4}$$

If interest lies only in a functional of the copula, instead of in its entire structure, we use Algorithm 2 to produce an approximation of the posterior distribution of the functional itself $\varphi = T(\boldsymbol{F})$.

It is important to note that the approximation might hold only asymptotically: for example, if the sample version of the Spearman's $\rho$ is used to approximate the posterior distribution of $\rho$, one has to consider that the sample version is only asymptotically unbiased. One advantage of the proposed Algorithm is that prior information is only provided for the marginal distributions and for $\varphi$; so the prior elicitation is easier: it is not necessary to define a prior distribution for the entire copula function.

Moreover, the method is robust with respect to different prior opinions about nonessential aspects of the dependence structure and with respect to the copula definition. The most important disadvantage of the method is its inefficiency when compared to a parametric copula, as usual in nonparametric or semiparametric setting; however this is true only under the assumption that the parametric copula is the true model.

From a computational perspective Algorithm 2 is quite demanding, since one needs to run a $BC_{EL}$ algorithm for each row of the posterior sample from the marginals. Even though the estimation of the marginal densities of the $X_{[\cdot,j]}$'s might not

```
[1:] For $b = 1, \dots, B$, use the $s$-th row of the posterior simulation $\lambda_1^{(b)}, \lambda_2^{(b)}, \dots, \lambda_d^{(b)}$ to
create a matrix of uniformly distributed *pseudo*-observations
    $\boldsymbol{u}^{(b)} = \begin{pmatrix} u_{11}^{(b)} & u_{12}^{(b)} & \dots & u_{1d}^{(b)} \\ u_{21}^{(b)} & u_{22}^{(b)} & \dots & u_{2d}^{(b)} \\ \dots & \dots & u_{ij}^{(b)} & \dots \\ u_{n1}^{(b)} & u_{n2}^{(b)} & \dots & u_{nd}^{(b)} \end{pmatrix}$
with $u_{ij}^{(b)} = F_j\left(x_{ij}; \lambda_j^{(b)}\right)$.
[2:] Given a prior distribution $\pi(\varphi)$ for the quantity of interest $\varphi$,
for $m = 1, \dots, M$,
    1. draw $\varphi^{(m)} \sim \pi(\varphi)$;
    2. compute $EL\left(\varphi^{(m)}; \boldsymbol{u}^{(b)}\right) = \omega_{mb}; b = 1, \dots, B$.
    3. take the average weight $\boldsymbol{\omega}_m = B^{-1} \sum_{b=1}^{B} \omega_{mb}$
end for
[3:] re-sample - with replacement - from $\{\left(\varphi^{(b)}, \boldsymbol{\omega}_b\right), b = 1, \dots, B\}$.
```

**Algorithm 2:** ABCOP algorithm

require a huge number of iterations $B$, still it might be very expensive to run $B$ different $BC_{EL}$ algorithms. To avoid this computational burden, we propose to modify the above algorithm by simply performing a single run of the $BC_{EL}$ algorithm, where, for each iteration $m = 1, \dots, M$, a randomly selected (among the $B$ rows) row $\boldsymbol{\lambda}^b$ is used to transform the actual data into pseudo-observations lying in $[0, 1]^d$.

## 4 An Example: Spearman's $\rho$

The definition of the Spearman's $\rho$ given in (4) can be interpreted as an average distance between the copula $C$ and the independence copula $\Pi(u, v) = uv$. Thus, in a $d$-dimensional setting the multivariate $\rho$ becomes

$$\begin{aligned} \rho &= \frac{\int_{[0,1]^d} C(\mathbf{u}) d\mathbf{u} - \int_{[0,1]^d} \Pi(\mathbf{u}) d\mathbf{u}}{\int_{[0,1]^d} M(\mathbf{u}) d\mathbf{u} - \int_{[0,1]^d} \Pi(\mathbf{u}) d\mathbf{u}} \\ &= \frac{d+1}{2^d - (d+1)} \left\{ 2^d \int_{[0,1]^d} C(\mathbf{u}) d\mathbf{u} - 1 \right\}. \end{aligned} \tag{5}$$

The multivariate extension of the empirical copula is

$$\hat{C}_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \mathbb{I}_{\left\{\hat{U}_{ijn} \le u_i\right\}} \qquad \mathit{for}\ \mathbf{u} = (u_1, u_2, \dots, u_n) \in [0, 1]^d \tag{6}$$

where $\hat{U}_{ijn} = \hat{F}(X_{ij})$ for $i = 1, \ldots, d$ and $\hat{F}(\cdot)$ is the empirical marginal distribution function. Therefore, a nonparametric estimator of $\rho$ is

$$\hat{\rho}_{1n} = h(d)\left\{2^d \int_{[0,1]^d} \hat{C}_n(\mathbf{u})d\mathbf{u} - 1\right\} = h(d)\left\{\frac{2^d}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}(1-\hat{U}_{ijn}) - 1\right\} \tag{7}$$

where $h(d) = (d+1)/(2^d - d - 1)$. An alternative estimator is

$$\hat{\rho}_{2n} = h(d)\left\{2^d \int_{[0,1]^d} \Pi(\mathbf{u})d\hat{C}(\mathbf{u}) - 1\right\} = h(d)\left\{\frac{2^d}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\hat{U}_{ijn} - 1\right\} \tag{8}$$

Asymptotic properties of these estimators are assessed in [6].

Once an estimator of the multivariate version of $\rho$ is available, it is possible to apply the procedure presented in Sect. 3. On the other hand, the variance of the proposed estimators can be explicitly computed only in few cases, for example in the case of the independence copula. [6] proposes to estimate it in a nonparametric way via a bootstrap method. Nevertheless, in practice this method tends to underestimate the variance, as it is shown in Fig. 1, where the frequentist procedure for a fixed $n$

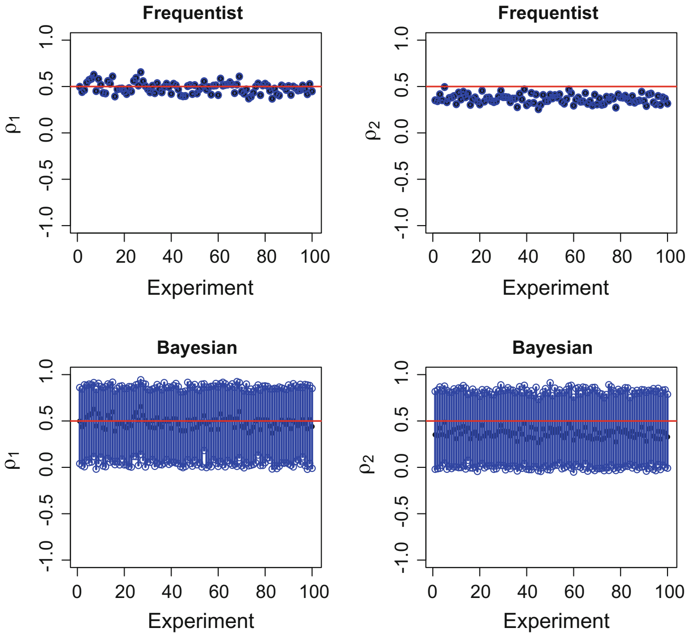


**Fig. 1** 100 simulations from a Clayton copula: sample size is 100; the true value of $\rho$ is equal to 0.5 (*red line*). The results for the frequentist procedure are available above, the ones for the Bayesian procedure are available below. The *black lines* are the point estimates of $\rho_1$, the *blue lines* represent the lower and the upper bounds of the intervals of level 0.95

leads to a coverage of about 10 % (coverage of 0 % for the interval of $\hat{\rho}_2$), while the proposed Bayesian method has the expected coverage, even if the average length is necessarily greater, about 0.822, i.e. the intervals contain almost half of the parameter space.

## 5 Further Research

Algorithm 2 produces an approximation of the posterior distribution of any summary of the multivariate dependence, once a multivariate estimator is available, as in the case of the Spearman's $\rho$. In some cases the analysis may be focused on measures of dependence as functions of some available conditioning variables. In the case of two response variables $X_1$ and $X_2$, both depending on the same covariate $Z$, the observations $(x_{1i}, x_{2i}, z_i)$ follow a distribution $F_{X_1,X_2|Z}(\cdot|z)$. [7] proposes the following estimator for the Spearman's $\rho$.

$$\hat{\rho}_n(x) = 12 \sum_{i=1}^{n} w_{ni}(x, h_n)(1 - \hat{U}_{i1})(1 - \hat{U}_{i2}) \tag{9}$$

where $\hat{U}_{i,j} = \sum_{i'=1}^{n} w_{i'}(x, h_n)\mathbb{I}(U_{i'j} \leq u_{ij})$ for $j = 1, 2$, $U_{ij} = F_j(x_{ij})$ and $w_{ij}(x, h_n)$ are appropriately chosen weights depending on $x_{ij}$ and a bandwidth $h_n$, for example kernel-based weights as the Nadaraya-Watson. Unfortunately, estimator (9) is based on an estimator of the conditional copula, given in [7], which is biased. A first simulation study implemented for 10,000 simulations of the function $\rho(x)$ (see Fig. 2) shows that, while the estimator (9) is not able to capture the true function

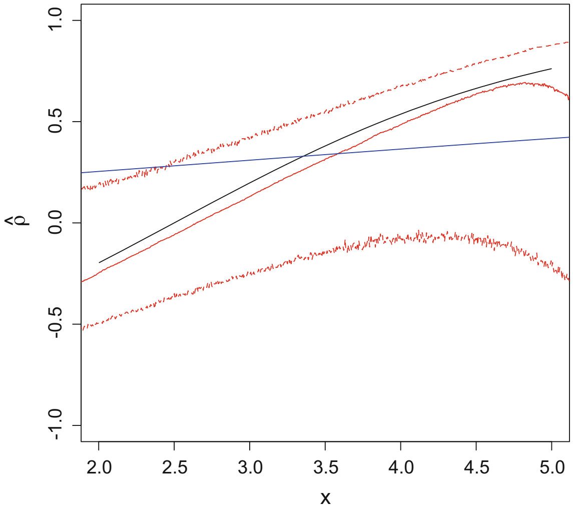


**Fig. 2** Simulations from the conditional Clayton copula based on 10,000 ABC simulations of $\rho(x)$ and 100 data points: true function $\rho(x)$ in *black*, Bayesian estimates in *red* (median, 0.05 and 0.95 credible bands), frequentist estimate in blue

(it underestimates the dependence among values), the Bayesian estimate obtained via Algorithm 2 can recover it, even if the variance increases as the value of the covariate increases. Further research will be focused on trying to understand why this happens and on producing more stable estimates.

## References

1. Sklar M (1959) Fonctions de répartition à *n* dimensions et leurs marges. Publ Inst Statist Univ Paris 8:229–231
2. Craiu VR, Sabeti A (2012) In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. J Multivar Anal 110:106–120
3. Mengersen K, Pudlo P, Robert CP (2013) Bayesian computation via empirical likelihood. Proc Natl Acad Sci 110(4):1321–1326
4. Min A, Czado C (2010) Bayesian inference for multivariate copulas using pair-copula constructions. J Financ Econometrics 8(4):511–546
5. Owen AB (2010) Empirical likelihood. Chapman & Hall/CRC Press, New York
6. Schmid F, Schmidt R (2007) Multivariate extensions of Spearmans Rho and related statistics. Stat Probab Lett 77(4):407–416
7. Gijbels I, Veraverbeke N, Omelka M (2011) Conditional copulas, association measures and their applications. Comput Stat Data Anal 55(5):1919–1932

# The Sign Test for Interval-Valued Data

**Przemysław Grzegorzewski and Martyna Śpiewak**

**Abstract** Two versions of the generalized sign test for interval-valued data are proposed. Each version correspond to a different view on the interval outcomes of the experiment—either the epistemic or the ontic one. As it is shown, each view yield different approaches to data analysis and statistical inference.

## 1 Introduction

Interval-valued data have drawn an increasing interest in recent years. However, a closed interval may be used to model two different types of information: the imprecise description of a point-valued quantity or the precise description of a set-valued entity.

Quite often the results of an experiment are imprecisely observed or are so uncertain that they are recorded as intervals containing the precise outcomes. Sometimes the exact value of a variable is hidden deliberately for some confidentiality reasons (see [7]). In all such cases intervals are considered as disjunctive sets representing incomplete information (*epistemic view*, according to [1]). In other words, an *epistemic set* $A$ contains an ill-known actual value of a point-valued quantity $x$, so we can write $x \in A$. Since it represents the epistemic state of an agent, it does not exist per se.

There are also situations when the experimental data appear as essentially interval-valued data describing a precise information (ranges of fluctuations of some physical measurements, time interval spanned by an activity, etc.). Such intervals are called conjunctive and correspond to the *ontic view* (see [1]). Thus an *ontic set* is the precise

P. Grzegorzewski (✉)
Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw, Poland
e-mail: pgrzeg@ibspan.waw.pl

P. Grzegorzewski · M. Śpiewak
Faculty of Mathematics and Information Science, Warsaw University of Technology,
Koszykowa 75, 00-662 Warsaw, Poland
e-mail: spiewakm2@student.mini.pw.edu.pl

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_34

representation of an objective entity, i.e. $A$ is a value of a set-valued variable $X$, so we can write $X = A$.

In this paper we suggest how to generalize the well-known sign test for the interval-valued data perceived from these two perspectives. We have chosen a distribution-free test deliberately to avoid problems in verifying assumptions on the underlying distribution. Indeed, yet we do not have satisfactory goodness-of-fit techniques for interval-valued data.

The paper is organized as follows: In Sect. 2 we recall the classical sign test. In Sect. 3 we introduce basic notations and concepts related to interval-valued data. Next, we propose two generalizations of the sign test adequate to each type of data: for epistemic sets in Sect. 4 and for ontic sets in Sect. 5.

## 2 The Sign Test

Many classical tests were derived assuming that samples come from the normal population. If we cannot warrant normality and a sample size is not large enough to perform an asymptotic test, nonparametric tests are usually recommended. For example, the popular z-test or t-test for the mean, which require normality, we may substitute by a very simple but useful sign test.

Although the sign test is not very powerful its most important advantage is the almost complete lack of assumptions on the population distribution. It does not also require a big sample. In this test the hypotheses concerns the median, not the mean. Both the mean and the median are good measures of central tendency and they coincide for symmetric distributions, but in any population the median always exists, which is not true for the mean, and the median is more robust to outlier as an estimate of location than the mean.

Suppose a random sample of $n$ independent observations $X_1, \ldots, X_n$ is drawn from the population with unknown median $M$. The only assumption is that the population distribution is continuous and strictly increasing in the vicinity of $M$. We verify the null hypothesis $H_0 : M = M_0$ with a corresponding one-sided or two-sided alternative. The idea of the sign test is very simple: if the data are consistent with the hypothesized median $M_0$ on the average half of the observations should lie above $M_0$ and a half below. Conversely, a significant disproportion between the number of positive signs of differences $X_i - M_0$ and the number of negative signs would lead to rejection of $H_0$. The test statistic delivers the number of observed "plus" signs and is defined as follows

$$T = \sum_{i=1}^{n} \mathbb{I}(X_i - M_0 > 0), \tag{1}$$

where $\mathbb{I}(\rho) = 1$ if a sentence $\rho$ is true and $\mathbb{I}(\rho) = 0$ otherwise.

The sampling distribution of $T$ is binomial $\text{Bin}(n, \theta)$ with parameters $n$ and $\theta$ which is equal to 0.5 if $H_0$ holds. For a one-sided upper-tailed alternative $H_1 : M > M_0$ we reject $H_0$ if $T \geqslant k_\alpha$, where $k_\alpha$ is chosen to be the smallest integer which satisfies $\mathbb{P}(T \geqslant k_\alpha | H_0) = \sum_{i=k_\alpha}^{n} \binom{n}{i} 0.5^n \leqslant \alpha$ and $\alpha$ is an accepted significance level. Similarly, for a one-sided lower-tailed alternative $H_1 : M < M_0$ we reject $H_0$ if $T \leqslant k'_\alpha$, where $k'_\alpha = n - k_\alpha$. And finally, for a two-sided alternative $H_1 : M \neq M_0$ we reject $H_0$ if $T \geqslant k_{\alpha/2}$ or $T \leqslant k'_{\alpha/2}$.

The sign test is also applicable to paired-sample data. Suppose we have two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ representing “pretreatment” and “posttreatment” observations on each of $n$ subjects (patients, blocks, etc.), respectively. Then we consider the null hypothesis that the treatment effect is not significant, i.e. $H_0 : M(X - Y) = 0$, where $M(X - Y)$ stands for the median of the difference between the pretreatment and the posttreatment. In our paired-sample problem the test statistic is given by

$$T = \sum_{i=1}^{n} \mathbb{I}(X_i - Y_i > 0), \tag{2}$$

while the rejection criteria remain as for the one-sample problem.

A generalization of the sign test for fuzzy data was proposed by Grzegorzewski [3, 5]. Below we suggest how to generalize the sign test for interval-valued data.

## 3 Interval-Valued Data

Let $\mathcal{K}_c(\mathbb{R}) = \{[u, v] : u, v \in \mathbb{R},\ u \leqslant v\}$ denote the family of all non-empty closed and bounded intervals in the real line $\mathbb{R}$. Each compact interval $A \in \mathcal{K}_c(\mathbb{R})$ can be expressed by its endpoints, i.e. $A = [\underline{a}, \overline{a}]$. Alternatively, the notation $A = [\text{mid}\, A \pm \text{spr}\, A]$, with $\text{spr}\, A \geqslant 0$, where $\text{mid}\, A = \frac{1}{2}(\underline{a} + \overline{a})$ is the mid-point (center) of the interval $A$ and $\text{spr}\, A = \frac{1}{2}(\overline{a} - \underline{a})$ is the spread (radius) of $A$, can be considered.

To handle intervals a natural arithmetic on $\mathcal{K}_c(\mathbb{R})$ is defined by means of the Minkowski addition and the product by scalars, given by

$$A + B = \{a + b : a \in A, b \in B\} \quad \text{and} \quad \lambda A = \{\lambda a : a \in A\},$$

for any $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. These two operations can be jointly expressed in terms of the mid/spr representation of the intervals as $A + \lambda B = [(\text{mid}\, A + \lambda \text{mid}\, B) \pm (\text{spr}\, A + |\lambda| \text{spr}\, B)]$, while using the endpoints of the intervals we obtain $A + B = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]$, $A - B = [\underline{a} - \overline{b}, \overline{a} - \underline{b}]$ and $\lambda A = [\min\{\lambda \underline{a}, \lambda \overline{a}\}, \max\{\lambda \underline{a}, \lambda \overline{a}\}]$.

It should be noted that the space $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear but semi linear, due to the lack of the opposite element with respect to the Minkowski addition: in general, $A + (-1)A \neq \{0\}$, unless $A = \{a\}$ is a singleton.

Although we use the same notation and basic operations on intervals both for the epistemic and ontic approach, there are significant differences in statistics of interval-valued data perceived from those two perspectives. In the epistemic approach we deal with usual random variables which attribute to each random event a real value, only its perception is not known precisely but exact to interval. On the other hand, in the ontic approach we deal with random intervals defined as follows.

**Definition 1** Given a probability space $(\Omega, \mathcal{A}, P)$, a mapping $X : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ is said to be a *random interval* (interval-valued random set) if it is Borel-measurable with the Borel $\sigma$-field generated by the topology associated with by the Hausdorff metric on $\mathcal{K}_c(\mathbb{R})$.

Equivalently, a mapping $X : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ is a random interval if $\operatorname{mid} X : \Omega \to \mathbb{R}$ and $\operatorname{spr} X : \Omega \to \mathbb{R}_+ \cup \{0\}$ are random variables defined as the mid-point and the spread of the interval $X(\omega)$, respectively, for each $\omega \in \Omega$.

## 4 The Sign Test in the Epistemic Perspective

Within the epistemic view let us consider a sequence of interval observations $X_1 = [\underline{x}_1, \overline{x}_1], \dots, X_n = [\underline{x}_n, \overline{x}_n]$, which are perceptions of the unknown true outcomes $x_1, \dots, x_n$ of the experiment, where $x_i \in X_i$. As in the classical case we assume that our observations come from the unknown distribution with a median $M$ and our goal is to verify a hypothesis $H_0 : M = M_0$ against the alternative $H_1 : M > M_0$. Suppose $T$ denotes test statistic (1). Our goal now is to find a set of possible values that the test statistics can assume, i.e.

$$T_I = \{T(x_1, \dots, x_n) : x_i \in X_i\}. \tag{3}$$

In general finding the set which is guaranteed to contain the actual range of statistic which may be assumed for any possible real values belonging to intervals that form interval-valued data is not easy. Moreover, in some cases it is even impossible in a reasonable time (e.g. determining the sample variance for arbitrary sample of the interval data perceived from the epistemic perspective is the NP-hard problem, see [9]). Fortunately, in the case of the sign test statistic to find its enclosure is not only possible but even easy. In fact, it will be enough to identify situations when (3) assumes its smallest value $\underline{t}$ and largest value $\overline{t}$.

Let us consider the following three situations. Firstly, if $\overline{x}_i < M_0$ then $\mathbb{I}(x_i - M_0 > 0) = 0$ for any $x_i \in [\underline{x}_i, \overline{x}_i]$, which means that in this case one may choose arbitrary $x_i \in [\underline{x}_i, \overline{x}_i]$ for computing both the upper or the lower bound of $T$. Secondly, if $\underline{x}_i < M_0 < \overline{x}_i$ then for designing $\underline{t}$ we may choose arbitrary $x_i \in [\underline{x}_i, M_0)$, for which we get $\mathbb{I}(x_i - M_0 > 0) = 0$. Similarly, to get $\overline{t}$ we may choose arbitrary $x_i \in (M_0, \overline{x}_i]$, for which $\mathbb{I}(x_i - M_0 > 0) = 1$. Finally, if $M_0 < \underline{x}_i$ then $\mathbb{I}(x_i - M_0 > 0) = 1$ for any $x_i \in [\underline{x}_i, \overline{x}_i]$ which means that one may choose arbitrary $x_i \in [\underline{x}_i, \overline{x}_i]$

for computing both the upper or the lower bound of $T$. As a conclusion we obtain both desired bounds of possible values of the test statistic

$$\underline{t} = \min\{T(x_1,\ldots,x_n) : x_i \in X_i\} = \sum_{i=1}^{n} \mathbb{I}(\underline{x}_i - M_0 > 0), \tag{4}$$

$$\overline{t} = \max\{T(x_1,\ldots,x_n) : x_i \in X_i\} = \sum_{i=1}^{n} \mathbb{I}(\overline{x}_i - M_0 > 0). \tag{5}$$

Since now the test statistic is no longer represented by a single value but by a bounded set with bounds (4) and (5), the corresponding p-value, required for making a decision, is not also a single value but form a set

$$p_I = \{\mathbb{P}(T \geqslant t \mid H_0) : t \in T_I\} \tag{6}$$

with the following bounds

$$\underline{p} = \min\{\mathbb{P}(T \geqslant t \mid H_0) : t \in T_I\} = \mathbb{P}(T \geqslant \overline{t} \mid H_0) = \sum_{i=\overline{t}}^{n} \binom{n}{i}\left(\frac{1}{2}\right)^n, \tag{7}$$

$$\overline{p} = \max\{\mathbb{P}(T \geqslant t \mid H_0) : t \in T_I\} = \mathbb{P}(T \geqslant \underline{t} \mid H_0) = \sum_{i=\underline{t}}^{n} \binom{n}{i}\left(\frac{1}{2}\right)^n. \tag{8}$$

It is worth noticing that $T_I \subseteq \{\underline{t}, \underline{t}+1, \ldots, \overline{t}\}$ and $p_I \subseteq \{\underline{p}, \underline{p}+1, \ldots, \overline{p}\}$.

In classical statistics we reject $H_0$ if a p-value $p$ is small enough, e.g. if $p < \alpha$, where $\alpha$ is the assumed significance level (typically $\alpha = 0.05$) and do not reject $H_0$ (accept $H_0$) otherwise. Unfortunately, in this case of the set-valued p-value the relation $p_I < \alpha$ means nothing. However, we may apply there the following natural algorithm proposed by Filzmoser and Viertl [2] to handle a fuzzy p-value: if $\overline{p} < \alpha$ then reject $H_0$; if $\alpha < \underline{p}$ then accept $H_0$; otherwise (i.e. if $\underline{p} \leqslant \alpha \leqslant \overline{p}$) we suspend the decision and, e.g., demand more observations to make a well-based decision.

It seems that the algorithm by Filzmoser and Viertl is well-grounded and may be recommended to practitioners. However, if one requires just one of the binary decisions—either reject or accept $H_0$—he/she may apply an appropriate randomization (see [4]).

By the similar reasoning we may generalize the sign test for two-sample paired data given by interval-valued observations $X_1 = [\underline{x}_1, \overline{x}_1], \ldots, X_n = [\underline{x}_n, \overline{x}_n]$ and $Y_1 = [\underline{y}_1, \overline{y}_1], \ldots, Y_n = [\underline{y}_n, \overline{y}_n]$. In this case the desired bounds of possible values of the test statistic are given as follows: $\underline{t} = \min\{T(x_1,\ldots,x_n,y_1,\ldots,y_n) : x_i \in X_i, y_i \in Y_i\} = \sum_{i=1}^{n} \mathbb{I}(\underline{x}_i > \overline{y}_i)$ and $\overline{t} = \max\{T(x_1,\ldots,x_n,y_1,\ldots,y_n) : x_i \in X_i, y_i \in Y_i\} = \sum_{i=1}^{n} \mathbb{I}(\overline{x}_i > \underline{y}_i)$.

## 5 The Sign Test in the Ontic Perspective

Now let us consider a sample of random intervals $X_1, \ldots, X_n$. Suppose we want to test hypothesis about the central tendency interval $M_0 = [\text{mid}\, M_0 \pm \text{spr}\, M_0]$ which is somehow "typical" for the population represented by our sample. However, firstly, one has to specify how to imagine a "typical" interval. Secondly, the relation between the object assumed to characterize the population and the true one has to be defined. And finally, the desired testing procedure has to be constructed. A test for interval-valued mean was proposed by Montenegro et al. [8]. Quite different approach was suggested by Ramos-Guajardo et al. [10, 11] who proposed a test for a hypothesis about a similarity between the expected value of a random interval and a fixed interval. In both cases the crucial difficulty is to find the distribution of the test statistic so the advised way-out is to use a bootstrap or to apply an asymptotic approach, provided a sample is large enough. Below we suggest a simple generalization of the sign test that avoids the above mentioned problem.

Let $Med = [\text{mid}\, Med \pm \text{spr}\, Med]$ denote the unknown interval-valued population median (see [12]). Assume that we want to test a hypothesis $H_0 : Med = M_0$, where $M_0 = [\text{mid}\, M_0 \pm \text{spr}\, M_0]$ is a fixed interval. Therefore, as a null hypothesis on location we will consider the following statement

$$H_0 : (\text{mid}\, Med = \text{mid}\, M_0 \text{ and } \text{spr}\, Med = \text{spr}\, M_0), \tag{9}$$

against the "two-sided" alternative $H_1 : \neg H_0$ that at least one of the equalities in (9) fails.

Since, according to Sect. 3, each interval-valued observation is completely described by its mid-point and spread, it seems obvious that if the null hypothesis (9) holds then the mid-point and the spread of $M_0$ should be "close" to the mid-points and spreads of the observed intervals, respectively.

Let us define the following two statistics:

$$T_1 = \sum_{i=1}^{n} \mathbb{I}(\text{mid}\, X_i - \text{mid}\, M_0 > 0), \tag{10}$$

$$T_2 = \sum_{i=1}^{n} \mathbb{I}(\text{spr}\, X_i - \text{spr}\, M_0 > 0). \tag{11}$$

This way our sign test for interval-valued data consists of two usual sign tests: one for the mid-points and the second for spreads. We reject the null hypothesis (9) if at least one of these tests indicates rejection, i.e. if either the median of $(\text{mid}\, X_1, \ldots, \text{mid}\, X_n)$ differs too much from $\text{mid}\, M_0$ or the median of $(\text{spr}\, X_1, \ldots, \text{spr}\, X_n)$ differs too much from $\text{spr}\, M_0$.

Let $\alpha \in (0, 1)$ denote the significance level of our generalized sign test. Then we have

$$\mathbb{P}(T_1 \in \mathcal{W}_1 \text{ or } T_2 \in \mathcal{W}_2 \mid H_0) \leqslant \alpha, \tag{12}$$

where $\mathcal{W}_1$ and $\mathcal{W}_2$ denote critical regions for our two subtests, i.e. for the mid-points and spreads, respectively. If the null hypothesis (9) holds then both $T_1$ and $T_2$ are binomially distributed, i.e. $\mathrm{Bin}(n, 0.5)$. However, a natural question that arises now is: How to find $\mathcal{W}_1$ and $\mathcal{W}_2$?

The left side of (12) equals $\mathbb{P}(T_1 \in \mathcal{W}_1 | H_0) + \mathbb{P}(T_2 \in \mathcal{W}_2 | H_0) - \mathbb{P}(T_1 \in \mathcal{W}_1, T_2 \in \mathcal{W}_2 | H_0)$. If we additionally assume that $T_1$ and $T_2$ are independent, which seems to be quite natural, then we get

$$\mathbb{P}(T_1 \in \mathcal{W}_1 | H_0) + \mathbb{P}(T_2 \in \mathcal{W}_2 | H_0) - \mathbb{P}(T_1 \in \mathcal{W}_1 | H_0)\mathbb{P}(T_2 \in \mathcal{W}_2 | H_0) \leqslant \alpha. \quad (13)$$

Let us introduce the following notation: $\alpha_1 = \mathbb{P}(T_1 \in \mathcal{W}_1 | H_0)$ and $\alpha_2 = \mathbb{P}(T_2 \in \mathcal{W}_2 | H_0)$. Then (13) can be expressed as follows

$$\alpha_1 + \alpha_2 - \alpha_1\alpha_2 \leqslant \alpha. \quad (14)$$

If we additionally assume that the closeness in mid-points is equally important as the closeness in spreads then $\alpha_1 = \alpha_2$ and hence (14) reduces to $\alpha_1(2 - \alpha_1) \leqslant \alpha$. Keeping in mind that $\alpha_1 \in (0, 1)$ we obtain the following desired relationship between $\alpha_1$ and $\alpha$

$$\alpha_1 \simeq 1 - \sqrt{1 - \alpha}. \quad (15)$$

Therefore, going back to Sect. 2, we obtain the critical region $\mathcal{W}_1 = \mathcal{W}_2 = [0, k'_{\alpha_1/2}] \cup [k_{\alpha_1/2}, n]$, where $k_{\alpha_1}$ is chosen to be the smallest integer which satisfies $\sum_{i=k_{\alpha_1/2}}^{n} \binom{n}{i} 0.5^n \leqslant \frac{\alpha_1}{2}$, $k'_{\alpha_1/2} = n - k_{\alpha_1/2}$ and $\alpha_1$ is given by (15).

It is worth noting that in a case when no identical importance is connected with the mid-points and spreads one may consider different $\mathcal{W}_1$ and $\mathcal{W}_2$ corresponding to different $\alpha_1$ and $\alpha_2$, respectively, which satisfy (14).

Unfortunately, the sign test designed for random intervals cannot be directly applied for the "one-sided" alternatives because intervals are not linearly ordered and for two intervals $M$ and $M_0$ the meaning of the expression like $M$ "is greater than" $M_0$ is neither clear nor obvious.

We may also calculate a p-value of the generalized sign test. If $t_1$ and $t_2$ are the observed values of statistics $T_1$ and $T_2$, respectively, then $p_1 = 2\min\{\mathbb{P}(T_1 \leqslant t_1 | H_0), \mathbb{P}(T_1 \geqslant t_1 | H_0)\}$ and $p_2 = 2\min\{\mathbb{P}(T_2 \leqslant t_2 | H_0), \mathbb{P}(T_2 \geqslant t_2 | H_0)\}$ are p-values of the two subtests oriented on the mid-points and spreads, respectively. Assuming independence of $T_1$ and $T_2$ and applying the same reasoning as used above, we obtain the p-value of the overall test

$$p = p_1 + p_2 - p_1 p_2. \quad (16)$$

Please note, that the p-value defined for the ontic approach by (16) is a real number from the unit interval, as in classical case and not as in the epistemic approach described in Sect. 4.

## 6 Conclusions

We have proposed two generalizations of the sign test designed for two different views on interval-valued data. However, one should be aware the distinction between ontic and epistemic sets because there is a risk of misusing even basic notions and tools. Both ontic and epistemic view yield different approaches to data analysis and statistical inference. Thus in the paper we have proposed two generalizations of the sign test designed for two different views on interval-valued data. Of course, many questions are still open. In particular, the statistical properties of both generalizations are of interest. Moreover, the case of dependent statistics $T_1$ and $T_2$ considered in the ontic perspective, as well as the one-sided test for random intervals would be also of interest.

## References

1. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: ontic vs. epistemic views. Int J approximate Reasoning 55:1502–1518
2. Filzmoser P, Viertl R (2004) Testing hypotheses with fuzzy data. The fuzzy p-value. Metrika 59:21–29
3. Grzegorzewski P (1998) Statistical inference about the median from vague data. Control Cybern 27:447–464
4. Grzegorzewski P (2001) Fuzzy tests—defuzzification and randomization. Fuzzy Sets Syst 118:437–446
5. Grzegorzewski P (2004) Distribution-free tests for vague data. In: Lopez-Diaz M et al (eds) Soft methodology and random information systems. Springer, Heidelberg, pp 495–502
6. Grzegorzewski P, Ramos-Guajardo AB (2015) Similarity based one-sided tests for the expected value and interval data. In: Proceedings of EUSFLAT 2015. Atlantis Press, pp 960–966
7. Kreinovich V, Servin C (2015) How to test hypotheses when exact values are replaced by intervals to protect privacy: case of t-tests. Departamental Technical Reports (CS). Paper 892. University of Texas at El Paso
8. Montenegro M, Casals MR, Colubi A, Gil MA (2008) Testing two-sided hypothesis about the mean of an interval-valued random set. In: Dubois D et al. (eds) Soft Methods for handling variability and imprecision. Springer, pp 133–139
9. Nguyen HT, Kreinovich V, Wu B, Xiang G (2012) Computing statistics under interval and fuzzy uncertainty. Springer
10. Ramos-Guajardo AB (2014) Similarity test for the expectation of a random interval and a fixed interval. In: Grzegorzewski P et al. (eds) Strengthening links between data analysis and soft computing. Springer, pp 175–182
11. Ramos-Guajardo AB, Colubi A, González-Rodríguez G (2014) Inclusion degree tests for the Aumann expectation of a random interval. Inf Sci 288:412–422
12. Sinova B, Casals MR, Colubi A, Gil AM (2010) The median of a random interval. In: Borgelt C et al. (eds) Combining soft computing and statistical methods in data analysis. Springer, pp 575–583

# Probability Distributions Related to Fuzzy P-Values

**Olgierd Hryniewicz**

**Abstract** In the paper we have considered different approaches for the calculation of the p-value for fuzzy statistical tests. For the particular problem of testing hypotheses about the mean in the normal distribution with known standard deviation, and a certain type of fuzziness (both in data and tested hypotheses) we have found probability distributions of the respective defuzzified p-values. These distributions let us evaluate the compatibility of the observed data with the assumed hypothetical model.

## 1 Introduction

The concept of p-value is probably the most frequently used and misused concept of statistics. It was formally introduced by R.A. Fisher in the 1920's, but practically it was used earlier, e.g., in works of Karl Pearson. Its usage represents an inductive approach to statistical data analysis. Many practitioners, trying to interpret results of their experiments, mix this approach with a deductive one introduced by Neyman and Pearson, and arrive at completely false conclusions. The situation is even more complicated when we take into account the third paradigm of data analysis—the Bayesian one. In this approach—in its "objective" version proposed by Jeffreys—an uninformative prior distribution defined on the set of all considered hypotheses is introduced, and then a posterior probability distribution, conditioned on the observed value of a certain test statistic, is calculated. The hypothesis with the highest value of this posterior probability is taken as the most plausible.

The controversies between different approaches to the interpretation of statistical tests are even amplified when we consider the problem of statistical testing in a fuzzy statistical environment. By a fuzzy statistical environment we understand situation when both statistical data and/or statistical hypotheses can be imprecisely perceived or defined. Statistical tests used in this environment are usually called fuzzy statistical tests. The problems with the understanding of fuzzy statistical tests begin with the used interpretation of the concept of a fuzzy random variable. Depending on

O. Hryniewicz (✉)
Systems Research Institute, Newelska 6, 01-447 Warsaw, Poland
e-mail: hryniewi@ibspan.waw.pl

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_35

"epistemic" or "ontic" interpretation of fuzzy random data (see [3] for more information) the interpretation of the results of fuzzy statistical tests may be quite different.

The paper is organized as follows. In the second section we present the concept of p-value, as it is used and interpreted in classical statistics. We begin with some historical remarks, then we present mathematical description of the concept, and finally present its often disputable interpretation. In the third section we present different approaches to the usage of the concept of p-value in fuzzy statistical tests. We discuss the application of a fuzzy p-value, and a crisp p-value that can be used in a fuzzy statistical environment. Using simulation methods we analyze differences between probability distributions of defuzzified versions of p-values. The paper is concluded in its last section.

## 2 The Concept of P-Value—A Crisp Case

The concept of p-value has been defined in many ways. Below, following [17], we present a general definition that can be used in further considerations. Let $\mathbf{X}$ be random data described by a continuous density function $f(\mathbf{x})$. Let us also assume that in our decision model this density is completely specified, and forms our hypothetical model $H_0$. Compatibility of this hypothetical model with observed data $\mathbf{x}$ is evaluated using a certain statistic $T(\mathbf{X})$ whose large values indicate less compatibility. The $p$ value is then defined as

$$p = P(T(\mathbf{X}) \geq T(\mathbf{x})). \tag{1}$$

Let us reformulate this definition making it more understandable, but in some cases more difficult for computation. Let $M$ be our hypothetical model (parameter of a distribution, cumulative probability function, etc.), and $M_x(X)$ be a sample statistic that describes $M$. Let $d(X) = d(M_x(X) - M)$ be a non-negative function that measures appropriately defined "distance" between $M_x(X)$ and $M$. Then, the definition (1) may be reformulated as

$$p = P(T^{'}(d(X)) \geq T^{'}(d(x))). \tag{2}$$

This representation is especially useful for testing hypotheses about a parameter of location. For example, in testing hypotheses about the expected value of a normally distributed random variable with known value of $\sigma$ such that $\sigma\sqrt{n} = 1$ we have the following simple formulae for the calculation of p-values [16]:

$$p_l = \Phi(\mu_0 - \bar{x}), \tag{3}$$

for testing the one-sided null hypothesis $H_0 : \mu \leq \mu_0$ against the alternative $H_A : \mu > \mu_0$,

$$p_u = \Phi(\bar{x} - \mu_0), \tag{4}$$

for testing the one-sided null hypothesis $H_0 : \mu \geq \mu_0$ against the alternative $H_A : \mu < \mu_0$, and

$$p_u = \Phi(-|\bar{x} - \mu_0|), \tag{5}$$

for testing the two-sided null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_A : \mu \neq \mu_0$, where $\Phi(.)$ is the cdf function of the standard normal distribution.

For the calculation of p-values we usually require that statistics $T$ (or $T^{'}$) should be pivotal, i.e., their probability distribution should not depend on unknown parameters. It always can be done if our model does not contain nuisance parameters. We need to know the cdf $F_T$ of the distribution of $T$ under the null hypothesis, and then the observed p-values $F_T(T(\mathbf{x}))$ are realizations of a uniformly distributed random variables.

The usage of the concept of p-value raises many problems and questions. Its calculation is simple in many practical cases (e.g., for the normal distribution). However, in the case of discrete distributions (non-parametric tests!), asymmetric null distribution of $T$, and the existence of nuisance parameters, even the methods applied in its computing are disputable (see, e.g., [1, 6]). However, the main problem, still unsolved despite dozens of papers which have been devoted to it, is with its interpretation. For example, this is by no means the probability that the tested hypothesis is true in the frequentist interpretation of probability. Selke et al. [17] found in simulation experiments, that the percentage of cases when the hypothesis is true, but the reported p-value is close to 0.05 (a typical critical value in decision making) is much higher than 0.05. Moreover, even if we interpret p-values as measures of support of respective hypotheses, this interpretation, as it was shown by Schervish [16], is logically incoherent. It has to be noted, however, then when we use other approaches to probability (Bayesian, fiducial) the usage of the word "probability" in the description of the p-value is justified. For example [16], p-values calculated for one-sided hypotheses are posterior probabilities (in the sense of Jeffreys) or fiducial probabilities (in the sense of Fisher).

## 3 The Concept of P-Value—A Fuzzy Case

The necessity of taking into account fuzzy imprecision while solving statistical problems has been shown in many publications. Interesting overviews can be found, e.g., in [7, 8]. The applicability of fuzzy statistics in solving practical problems has been also well described in many papers (in particular, for the application in the area of reliability and statistical quality control, see [9, 12]). Due to limited volume of this paper we do not want to repeat arguments that imprecise (fuzzy) data do exist in practice. However, especially in the context of this paper, we want to stress the necessity to consider fuzzy statistical hypotheses. It has been noticed by many authors that the concept of p-value does not work properly for large samples and point-wise statistical hypotheses. In such cases the reported p-value will be usually very small, indicating the rejection of the tested hypothesis. A good example was given by Hryniewicz [10] who considered the problem of testing statistical independence. As this concept is very precisely defined, even small, and unimportant from a practical point of view, departures from the ideal situation will lead to unnecessary rejection of the tested hypothesis. Another example is related to the existence of small correlations between sample observations. The existence of these correlations changes the distribution of the test statistic, usually influencing its variability. Therefore, the observed test error rates may differ from the assumed ones. These, and similar, examples let us convince that the consideration of "relaxed" (fuzzy) hypotheses makes practical sense.

The crucial, in statistical practice, concept of a fuzzy p-value was introduced independently in papers of Filzmoser and Viertl [5] and Denœux et al. [4]. Filzmoser and Viertl [5] considered, using somewhat different terminology, $\delta$-cuts (or $\delta$-level sets) of the fuzzy observed values $\tilde{x}$ of the test statistic $T$ defined by (1). Then, they calculated the probabilities by the computation of respective areas under the pdf function of $T$. For example, for one-sided tests, and imprecise data described by closed and finite intervals $[T_1(\delta), T_2(\delta)]$ the fuzzy p-values of these tests are described by the following sets of intervals

$$C_\delta^L(\tilde{p}) = [P(T \leq T_1(\delta)), P(T \leq T_2(\delta))], \delta \in (0, 1], \tag{6}$$

and

$$C_\delta^U(\tilde{p}) = [P(T \geq T_1(\delta)), P(T \geq T_2(\delta))], \delta \in (0, 1]. \tag{7}$$

For two-sided hypothesis the respective formulae are more complicated.

Denœux et al. [4] in their definition of the fuzzy p-value used a computationally more effective approach. They noticed that in many cases there exist closed formulae for the computation of p-values in crisp cases, such as (3)–(5) when the test statistic $T$ is at least asymptotically normally distributed. In such cases they directly apply Zadeh's extension principle arriving at the fuzzy p-value $\tilde{p}$ whose membership function is represented by the nested set of respective $\delta$-cuts.

What really differs these two proposals is the suggested method of decision making. In both cases the authors assume, as in the Neyman-Pearson methodology, that the null hypothesis $H_0$ is tested against the alternative $H_A$. Filzmoser and Viertl [5] propose a very restrictive approach. They assume a certain critical value $\alpha$ (e.g., equal to 0.05), and propose to reject (accept) $H_0$ only when the whole support of $\tilde{p}$ is situated to the left (right) of $\alpha$. Otherwise, the decision cannot be made. The procedure proposed by Denœux et al. is far less restrictive. They interpret the membership function of the fuzzy p-value as a possibility distribution. Then, they calculate possibilities: $\Pi_1 = \Pi(\tilde{p} \leq \alpha)$ and $\Pi_0 = \Pi(\tilde{p} > \alpha)$, and propose to reject the null hypothesis if $\Pi_1 > \Pi_0$, and accept, otherwise. Another, well-grounded in the theory of possibility and imprecise probabilities, approach was proposed by Couso and Sanchez [2]. They have shown how the fuzzy p-value can be interpreted in terms of a second order possibility measure. Then, they proposed a defuzzified representation of the fuzzy p-value by the following crisp interval $[\underline{p_{val}(\tilde{\mathbf{x}})}], \overline{p_{val}(\tilde{\mathbf{x}})}]$, where

$$\underline{p_{val}(\tilde{\mathbf{x}})} = \int_0^1 \inf[p_{val}(\tilde{\mathbf{x}})]_\delta d\delta, \tag{8}$$

$$\overline{p_{val}(\tilde{\mathbf{x}})} = \int_0^1 \sup[p_{val}(\tilde{\mathbf{x}})]_\delta d\delta. \tag{9}$$

Let us propose now to look at the fuzzy p-value from a different point of view. In this approach we will use the concept of testing interval hypotheses introduced by Lehmann [14]. Schervish [16] shows that this concept generalizes testing of both one-sided and two-sided hypotheses. For testing the interval hypothesis $\mu \in (\mu_1, \mu_2)$ about the mean in the normal distribution with the known value of $\sigma$ the formula for the respective (crisp) p-value is given as [16]

$$p_{\mu_1,\mu_2}(x) = \begin{cases} \Phi(x-\mu_1) + \Phi(x-\mu_2), & \text{if } x < 0.5(\mu_1+\mu_2) \\ \Phi(\mu_1-x) + \Phi(\mu_2-x), & \text{if } x \geq 0.5(\mu_1+\mu_2) \end{cases} \tag{10}$$

Suppose now that we observe fuzzy random data $\tilde{X} = X_0 + \tilde{W}$, where $X_0$ is a crisp random variable that represents the most plausible value of the unknown origin of the observed fuzzy random variable, and $\tilde{W}$ represents the fuzzy part of $\tilde{X}$ which is *independent* of $X_0$. Moreover, we assume that our hypothetical value $\mu$ may also be imprecise, and is represented by a fuzzy number $\tilde{\mu}$. Consider now a "distance" between an observed value of $\tilde{X}$, namely $\tilde{x}$, and $\tilde{\mu}$. For a given $\delta$-level, $\tilde{x}$ is represented by its $\delta$-cut $(x_{L,\delta}, x_{L,\delta})$, and $\tilde{\mu}$ by its $\delta$-cut $(\mu_{L,\delta}, \mu_{L,\delta})$. Then, by some simple operations on interval-valued numbers we can show that this difference is equivalent to the distance between the observed value $x_0$, and the interval $[\mu_{L,\delta} - x_{r,\delta}, \mu_{R,\delta} - x_{L,\delta}]$. Hence, on a given $\delta$-cut we can consider a fuzzy statistical test as a test of an interval hypothesis about $X_0$. Note, that this means that there is no difference in testing statistical hypotheses using fuzzy data and fuzzy hypothetical values. In both cases, on the given $\delta$-level, the test is described by a single number. For instance, for a fuzzy test about the expected value of the normal distribution (with known value of $\sigma$)

$$p_{int,\delta}(x) = \begin{cases} \Phi(x_0-u_{1,\delta}) + \Phi(x-u_{2,\delta}), & \text{if } x < 0.5(u_1+u_2) \\ \Phi(u_{1,\delta}-x_0) + \Phi(u_{2,\delta}-x_0), & \text{if } x \geq 0.5(u_1+u_2) \end{cases} \tag{11}$$

where $u_{1,\delta} = \mu_{L,\delta} - x_{r,\delta}$, and $u_{2,\delta} = \mu_{R,\delta} - x_{L,\delta}$. Intervals $(0, p_{int,\delta}(x))$ can be regarded as the representation of the fuzzy p-value $\tilde{p}_{int}$. This fuzzy value can be defuzzified using (9). Unfortunately, $\tilde{p}_{int}$ is not a fuzzy number, as its membership function is not convex. It is a consequence of the incoherence of p-values, as it was noted by Schervish [16].

Finally, let us consider a possibilistic approach to statistical testing of hypotheses in a fuzzy environment proposed by Hryniewicz [11]. He assumed that for a test on a given significance level $\alpha$ only those values of fuzzy data and fuzzy hypotheses should be taken into account whose values of membership functions are not smaller than $\alpha$. This assumption can be justified by treating $\alpha$ as a certain measure of possibility. To illustrate this approach let us consider the case of a two-sided hypothesis about the value $\theta$ of a parameter of a certain probability distribution. In presence of fuzzy data $\tilde{x}$ we can calculate, using a classical approach proposed by Kruse and Meyer [13], fuzzy confidence intervals for $\theta$. Denote by $\mu_X(x)$ and $\mu_\theta(\theta_0)$ the membership functions of fuzzy data $\tilde{x}$ and fuzzy hypothesis $\tilde{\mu}_0$, respectively. Let $x_0$ be such that $\mu_X(x_0) = 1$, $\theta_0$ be such that $\mu_\theta(\theta_0) = 1$, $[\mu_{0,L}^\delta, \mu_{0,R}^\delta]$ be the $\delta$-level set of the fuzzy hypothesis $\tilde{\mu}_0$, and $[C_{\alpha,L}^\delta, C_{\alpha,R}^\delta]$ be the $\delta$-level set of the fuzzy confidence interval of $\theta$, calculated on the confidence level $\beta = 1 - \alpha/2$. Then, the possibilistic p-value $p_{ps}$ for the two-sided test about $\theta$ is given by

$$p_{ps} = \begin{cases} \sup\limits_{\delta} C_{\delta,R} = \mu_{0,L}^\delta, & \text{if } x_0 < \mu_0 \\ \sup\limits_{\delta} C_{\delta,L} = \mu_{0,R}^\delta, & \text{otherwise} \end{cases} \tag{12}$$

Similar conditions can be formulated also for one-sided hypotheses about the values of parameters of probability distributions. The advantage of this approach stems from the fact that we do not need any defuzzification procedure. For both fuzzy data and fuzzy hypotheses the value of $p_{ps}$ is crisp, and thus seemingly easier for interpretation.

## 4 Probability Distributions of Deffuzified P-Values

When we want to compare different approaches for the calculation of p-values in fuzzy environment we need to define certain comparable characteristics. Unfortunately, even in the crisp case such characteristics do not exist. The analysis of the probability distribution of the p-value, when the model of data is different from the hypothetical one, seems to be one of a few options. We know that when the null hypothesis is true this distribution is uniform. In other cases a general answer about the distribution of the p-value seemingly does not exist. However, when we test a double-sided hypothesis about the expected value $\theta$ of a normally distributed test statistic with a known value of its standard deviation $\sigma$, the p-value is distributed according to the power law distribution $F(p) = p^\gamma$. Let $D = |\theta - \theta_0| = D_\sigma * \sigma$ be the shift of the considered expected value. Without the lost of generalization we assume that the hypothetical value $\theta_0$ is equal to zero, and let $z = D_\sigma/\sqrt{n}$. Then in a simulation experiment we have found a very precise ($R^2 = 0.997$) relationship for the calculation of an approximate value of $\gamma$

$$\gamma_0 = 0,056908n^2z^2 - 0,45953nz + 1 \tag{13}$$

In extensive simulation experiments we have tried to find similar relationships for lower and upper values of the p-values defuzzified according to (8)–(9), interval based p-values $p_{int}$ calculated according to (11), and defuzzified according to (9). Moreover we have considered possibilistic p-values $p_{ps}$ calculated according to (12). In all these cases using simulation experiments we have found good approximations of functions that link the value of $\gamma$ with characteristics of fuzzy data and fuzzy hypotheses. These approximations have a following general form $\gamma_{fpv} = \gamma_0 + \gamma_{f,fpv}$, where $\gamma_{f,fpf}$ is a part related to the fuzziness of data and/or hypotheses. In our experiments, whose results are described below, we have assumed that fuzzy data are described by randomly chosen triangular membership functions with constant support $s_x$. Our considered fuzzy hypotheses have membership functions symmetric around a precise null hypothesis with the support equal to $s_m$.

For the lower and upper limits of the fuzzy p-values, defuzzified according to (8)–(9), the respective formulae for the fuzzy part of $\gamma_{fpv}$ are the following

$$\begin{aligned}\gamma_{f,fL} &= -5,737s_x - 7,669s_m + 9,129s_x^2 + 56,587s_m^2 - 0,224ns_m + \\ &- 178,99zs_m + 75,686s_xs_m - 3,795s_x^3 + 0,0015n^2s_m + \\ &- 69,913s_x^2s_m + 694,29s_m^2z - 306,57s_m^2s_x + 271,70s_x^2s_m^2,\end{aligned} \tag{14}$$

$$\begin{aligned}\gamma_{f,fR} &= -83,52zs_x - 0,223ns_m - 272,1zs_m + 36,05s_xs_m + \\ &+ 1483,2z^2s_m + 0,101ns_x^2 - 110,54zs_x^2 + 0,0188n^2s_m^2 + 1492,3z^2s_x^2.\end{aligned} \tag{15}$$

The accuracy of the approximation for $\gamma_{f,fL}$ is good ($R^2 = 0,879$). However, for $\gamma_{f,fR}$ the accuracy looks only reasonable ($R^2 = 0,744$). The respective formulae for the cases of fuzzy interval hypotheses, and possibilistic p-value, are the following

$$\begin{aligned}\gamma_{f,fint} &= -25,11s_m + 250,0s_m^2 - 0,018ns_m - 114,42zs_x + \\ &- 43,24zs_m + 24,78s_xs_m - 621,4s_m^3 + 459,76z^2s_x + \\ &+ 0,0222s_x^2n + 60,88s_x^2z - 80,09s_m^2s_x,\end{aligned} \tag{16}$$

$$\begin{aligned}\gamma_{f,fps} &= 4,086s_x^2 + 147,62s_m^2 - 0,231ns_m - 111,65zs_x - 547,26zs_m + \\ &+ 20,03s_xs_m - 3,114s_x^3 - 717,95s_m^3 + 304,75z^2s_x + 2739,7z^2s_m + 85,74zs_x^2 + \\ &+ 0,0262ns_x^2 - 17,94s_x^2s_m + 1,15ns_m^2 + 2052,2zs_m^2 - 10497,3z^2s_m^2.\end{aligned} \tag{17}$$

The accuracy of these approximations is also reasonable, $R^2 = 0,758$ and $R^2 = 0,825$, respectively. It has to be noted, however, that for some combinations of input parameters these approximations do not work well (possible negative values of $\gamma$).

Knowing the description of fuzziness (both for data and hypotheses) we can compute the respective value of $\gamma$, and thus the probability of having the p-value not greater than the observed one. Therefore, we may have an idea whether the observed p-value is small enough to reject the null hypothesis against its alternative.

## 5 Conclusions

In the paper we have considered different approaches for the calculation of the p-value for fuzzy statistical tests. For the particular problem of testing hypotheses about the mean in the normal distribution with known standard deviation, and a certain type of fuzziness (both in data and tested hypotheses) we have found probability distributions of the respective defuzzified p-values. These distributions let us evaluate the compatibility of the observed data with the assumed hypothetical model. Further research is needed for other popular statistical tests (e.g., a fuzzy version of Student's $t$), and for different types of fuzziness.

## References

1. Bayarri MJ, Berger J (2000) J Amer Stat Assoc 95(1127–1142):1157–1170
2. Couso I, Sánchez L (2008) Deffuzification of fuzzy p-values. In: Dubois D et al (eds) Soft methods for handling variability and imprecision. Springer, Heidelberg, pp 126–132
3. Couso I, Dubois D, Sánchez L (2014) Random sets and random fuzzy sets as Ill-perceived random variables. Springer, Heidelberg
4. Denœux T, Masson MH, Hébert PA (2005) Fuz Sets and Syst 153:1–28
5. Filzmoser P, Viertl R (2004) Metrika 59:21–29
6. Gibbons JD, Pratt JW (1975) Amer Stat 29:20–25
7. Gil MA, Hryniewicz O (2009) Statistics with Imprecise Data. In: Meyers RE (ed) Encyclopedia of complexity and systems science. Springer, Heidelberg, pp 8679–8690
8. Grzegorzewski P, Hryniewicz O (1997) Mathware Soft Comp 4:203–217
9. Grzegorzewski P, Hryniewicz O (2001) Soft methods in hypotheses testing. In: Ruan D, Kacprzyk J, Fedrizzi M (eds) Soft computing for risk evaluation and management. Physica Verlag, Heidelberg and New York, pp 55–72
10. Hryniewicz O (2006) On testing fuzzy independence. In: Lawry J et al (eds) Soft methods for integrated uncertainty modeling. Springer, Berlin Heidelberg, pp 29–36
11. Hryniewicz O (2006) Fuz Sets Syst 157:2665–2673
12. Hryniewicz O (2008) Soft Comp 12:229–234
13. Kruse R, Meyer KD (1987) Statistics with vague data. Riedel, Dodrecht
14. Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, New York
15. Lehmann EL (1993) J Am Stat Assoc 88:1242–1249
16. Schervish MJ (1996) The Amer Statistician 50:203–206
17. Sellke T, Bayarri MJ, Berger JO (2001) The Amer Statistician 55:62–71

# Probabilistic Semantics and Pragmatics for the Language of Uncertainty

**Stefan Kaufmann**

**Abstract** The idea that the probability of a conditional is the corresponding conditional probability has led something of an embattled existence in philosophy and linguistics. Part of the reason for the reluctance to embrace it has to do with certain technical difficulties (especially triviality). Even though solutions to the triviality problem are known to exist, their widespread adoption is hindered by their narrow range of data coverage and unclear relationship to established frameworks for modeling the dynamics of belief and conversation. This paper considers the case of *Bernoulli models* and proposes steps towards broadening the coverage of their application.

## 1 Introduction

This paper is concerned with the interpretation of conditional (*if-then*) sentences in a probabilistic framework. I take as my starting point the idea that the probability of a conditional *if A then C* is the conditional probability of $C$, given $A$ (henceforth "the Thesis"). Its theoretical and empirical ramifications have been studied extensively by philosophers [1, 2, among many others] and psychologists [28–30]. A general consensus has emerged that despite certain counterexamples [15, 19, 23, 26], its theoretical and empirical appeal is sufficient to warrant detailed investigation.

Nonetheless, the Thesis still has something of an embattled status in Philosophy of Language and Natural Language Semantics. A major factor contributing to this is undoubtedly the fact that it cannot be straightforwardly unified with the view that conditionals denote propositions in the usual sense. This was first established by Lewis's famous *triviality results* [24, 25], which inspired a formidable tradition of further observations and generalizations.

S. Kaufmann (✉)
Department of Linguistics, University of Connecticut, 365 Fairfield Way,
Storrs, CT 06269-1145, USA
e-mail: stefan.kaufmann@uconn.edu

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_36

In this paper I assume familiarity with the issue of triviality; the interested reader is referred to the numerous excellent surveys and expositions (e.g. [3, 6, 10]) and references therein. What I focus on instead is the fact that it is possible to uphold the Thesis while avoiding triviality. Specifically, I shall focus on van Fraassen's *Bernoulli models* for the assignment of probabilities to simple and compounded conditionals [33].[1] This approach was put forward at roughly the same time as the triviality results themselves, but despite a number of subsequent elaborations and explorations [9, 17, 18, 32], no account of conditionals based on it has as-yet gained significant currency.

One reason for this reluctance may be the fact that in some cases the probabilities predicted for certain conditionals are counter-intuitive. I have argued elsewhere for some of these challenges that they call for fine-tuning rather than abandonment of the approach [15–17], and I suspect that solutions for further problems can also be found. In this paper I shall address another potential impediment in the way of the Bernoulli model towards the mainstream, *viz.* its relatively narrow range of application.

Stepping back, there are good reasons to think not only that the Bernoulli approach deserves closer investigation, but also that this is a good time to carry out such a program. One such reason is a confluence of results between this approach and the *coherence*-based framework for subjective probability, which originated with de Finetti's work. De Finetti's ideas influenced the development of the Bernoulli framework via Jeffrey and Stalnaker [13, 32]; specifically, Jeffrey's proposal to treat conditionals as random variables was inspired by [5]. More recently, the full extent of the affinity was clarified in Gilio and Sanfilippo's explorations [8, 9], which uncovered parallels not only in basic ideas but also in concrete results and predictions (e.g., concerning probabilities of compounds with conditional constituents).

## 2 Some Data and Observations

At its core, a conditional *if A, then C* states that $C$ holds *on the supposition* that $A$. This idea goes back at least as far as Ramsey [31]. It underlies the standard semantic analysis in linguistics, which assumes that all conditionals involve a modal operator (which may or may not be overtly expressed in the sentence) whose domain of quantification is restricted by the antecedent ([20–22], inter alia). In a probabilistic framework, the natural analog of this idea is that conditionals are interpreted by *conditioning* on their antecedent.

This basic idea is straightforward enough. It raises a number of theoretical and empirical questions, however, as soon as we consider a somewhat broader range of phenomena. I list two in the remainder of this section.

[1] Van Fraassen dubbed them "Stalnaker Bernoulli models." I avoid this label in deference to Robert Stalnaker's contention that it suggests more credit for him than he deserves (p.c.).

## 2.1 Compounds of Conditionals

Compounded and embedded conditionals are well-formed and attested (here and below, I use the symbol '>' in formal renderings):

(1) a. If this vase will crack if it is dropped on wood, it will shatter if it is dropped on marble. $(W > C) > (M > S)$
b. If she drew a prime number, it is even, and if she drew an odd number, it is prime. $(P > E) \wedge (O > P)$

Such sentences pose challenges for the Thesis. The standard Bayesian calculus does not provide a way to calculate their probabilities in accordance with the Thesis. In Lewis's words, conditional probabilites are "probabilities only in name" [24], not probabilities that some proposition is true. Thus there is no straightforward way to extend a probabilistic account of conditionals to embeddings containing them.

## 2.2 Unconditionals

Another problem concerns so-called *unconditional* sentences, which share with conditionals the overall antecedent-consequent architecture, but are set apart by the fact that their antecedents are *interrogative* clauses:

(2) a. *Whether* Mary comes *or not*, we will have fun.
b. *Whether* John *or* Mary comes, we will have fun.
c. *Whoever* comes, we will have fun.
d. *No matter who* comes, we will have fun.

Interrogative clauses are typically analyzed as denoting sets of propositions, rather than just propositions as their declarative counterparts do [11, 12, 14]. It is widely agreed that in an unconditional, the conditional operator distributes over the propositions in the denotation of the antecedent. Thus for instance, (2a) and (2b) are equivalent to (3a) and (3b); likewise for the remaining sentences. This gives us some idea of what a semantic analysis of these sentences ought to predict.

(3) a. If Mary comes we will have fun, and if Mary doesn't come we will have fun.
b. If John comes we will have fun, and if Mary comes we will have fun.

However, the standard probabilistic calculus does not even provide us with a means to conditionalize on sets of propositions.

## 3 Bernoulli Models

**Definition 1** (*Probability model*) A probability model for the language of propositional logic is a structure $\langle \Omega, \mathcal{F}, Pr, V \rangle$, where

a. $\Omega$ is a non-empty set (of possible worlds);
b. $\mathcal{F}$ is a $\sigma$-algebra of subsets of $\Omega$ (propositions);
c. $Pr$ is a probability measure on $\mathcal{F}$; and
d. $V$ is a valuation function mapping sentence-world pairs to truth values in $\{0, 1\}$, subject to the following constraints:

$$V(\neg\varphi)(\omega) = 1 - V(\varphi)(\omega)$$
$$V(\varphi \wedge \psi)(\omega) = V(\varphi)(\omega) \cdot V(\psi)(\omega)$$

Although this is not required by the definition, no harm is done if we assume for simplicity that $\Omega$ is countable and $\mathcal{F}$ is the powerset of $\Omega$. This ensures that the $\sigma$-algebra can be defined and that the denotations of all atomic sentences and truth-functional compounds thereof receive probabilities under $Pr$; otherwise this would have to be stipulated separately.

To be able to talk about the probability of a sentence, I define a function $P$ mapping sentences to the expectations of their truth values: $P(\varphi) = E[V(\varphi)]$. Clearly in the present case this means that $P(\varphi) = Pr(\{\omega | V(\varphi)(\omega) = 1\})$, that is, the probability that $\varphi$ is true.

Of course, the material conditional is not the intended rendering of our natural language *if-then* sentences. Nor is there a way in general to extend $V$ to conditionals in such a way that their probabilities equal the corresponding conditional probabilities for all probability distributions. This is the lesson from the triviality results.[2]

A Bernoulli model is an extension of a probability model, incorporating a simple intuition about the interpretation of conditionals $\varphi > \psi$: Perform a series of trials (independent and identically distributed, according to $Pr$) in which a world is chosen from $\Omega$ (with replacement), until you pick a world at which the antecedent $\varphi$ is true. Check whether the consequent $\psi$ is also true at the same world. Sentences receive truth values relative to sequences of such trials (hence the term "Bernoulli" model).

**Definition 2** (*Bernoulli model*) Given a probability model $\langle \Omega, \mathcal{F}, Pr, V \rangle$, the corresponding Bernoulli model is the structure $\left\langle \Omega^*, \mathcal{F}^*, Pr^*, V^* \right\rangle$, where

a. $\Omega^*$ is the set of all countable sequences of worlds in $\Omega$. Notation:
'$\omega^*[n]$' is the $n$-th world in $\omega^*$, $n \geq 1$;
'$\omega^*(n)$' is the "tail" of $\omega^*$ starting with the $n$-th world.
b. $\mathcal{F}^*$ is the set of all Cartesian products $X_1 \times \cdots \times X_n \times \Omega^*$ for $X_i \in \mathcal{F}$;
c. $Pr^*$ is a product measure on $\mathcal{F}^*$ defined as follows:

$$Pr^*(X_1 \times \cdots \times X_n \times \Omega^*) = Pr(X_1) \times \cdots \times Pr(X_n)$$

[2] Except, that is, for models with no more than two distinct propositions in the domain of the probability distribution. [24] called such models "trivial."

d. $V^*$ is a function from pairs of sentences and sequences in $\Omega^*$ to truth values, defined as follows:

$$\begin{aligned}
V^*(p)(\omega^*) &= V(p)(\omega^*)[1] \text{ for atomic } p\\
V^*(\neg\varphi)(\omega^*) &= 1 - V^*(\varphi)(\omega^*)\\
V^*(\varphi \wedge \psi)(\omega^*) &= V^*(\varphi)(\omega^*) \cdot V^*(\psi)(\omega^*)\\
V^*(\varphi > \psi)(\omega^*) &= V^*(\psi)(\omega^*(n)) \text{ for the least } n \text{ s.t. } V^*(\varphi)(\omega^*(n)) = 1
\end{aligned}$$

In the last clause of the definition of $V^*$, the rule for conditionals with antecedent $\varphi$ at $\omega^*$ calls for inspection of the longest "tail" of $\omega^*$ at which $\varphi$ is true. If there is no such tail, the value of any conditional with antecedent $\varphi$ is undefined. But whenever $\varphi$ has positive probability, the set of these "$\varphi$-less" sequences has zero probability, thus the probability that the sentence is true equals the conditional probability that it is true, given that its truth value is defined.

As before, I define a function $P^*$ mapping sentences to the expectations of their truth values under $V^*$. Now, in a Bernoulli model, the probability $P^*(\varphi > \psi)$ of a conditional is both the probability of a set of sequences (namely those at which the conditional is true) and the conditional probability $P^*(\psi|\varphi)$.[3] Moreover, if $\varphi$ and $\psi$ do not contain conditionals (and thus are in the domain of $V$ in the underlying probability model), then $P^*(\psi|\varphi)$ also equals $P(\psi|\varphi)$. The probabilities of more complex compounds involving conditionals can likewise be calculated in terms of the probabilities of their conditional-free constituents. For details, see [9, 17, 18].

## 4 Interrogative Antecedents

I first define a simple auxiliary device: For arbitrary sequences $\omega^*$ and sentences $\varphi$, let $\omega^*{\uparrow}\varphi$ be defined as follows:

(4) $$\omega^*{\uparrow}\varphi = \begin{cases} \omega^*(n) & \text{for the least } n \text{ s.t. } V^*(\varphi)(\omega^*(n)) = 1\\ \omega^* & \text{if there is no such } n \end{cases}$$

Thus $\omega^*{\uparrow}\varphi$ is the longest tail at which $\varphi$ is true, referred to in the definition of $V^*$ above, if such a longest tail exists. Otherwise $\omega^*{\uparrow}\varphi$ is just $\omega^*$.

Consider first the interpretation of conditionals with interrogative antecedents. Recall that, as I stated above in Sect. 2.2, interrogative clauses are analyzed as denoting sets of propositions. Extend the definition of the $\cdot{\uparrow}\cdot$-operator to sets $\Phi$ of propositions as follows:

(5) $$\omega^*{\uparrow}\Phi = \{\omega^*{\uparrow}\varphi \mid \varphi \in \Phi\}$$

[3] Note that $V^*(\varphi > \psi)$ is defined with probability 1 if $P^*(\varphi) > 0$, and undefined with probability 1 if $P^*(\varphi) = 0$. In the latter case, the expectation of the conditional's truth value is undefined, as is the probability $P^*(\psi|\varphi)$, at least when defined as the ratio $P^*(\varphi \wedge \psi)/P^*(\varphi)$. But see [18] for a definition of conditional probability in Bernoulli models which is defined for certain zero-probability propositions on which one might want to conditionalize. ([8, 9] also define the "prevision" for $\varphi > \psi$ in such a way that it includes the case that the prevision of $\varphi$ is zero.).

Thus $\omega^*{\uparrow}\Phi$ can result in multiple "active" alternatives for $\omega^*$. We can then define three different conditional operators as follows:

(6) a. $V^*(\Phi >_\forall \psi)(\omega^*) = 1$ iff $V^*(\psi)(\omega^{*\prime}) = 1$ for all $\omega^{*\prime} \in \omega^*{\uparrow}\Phi$
b. $V^*(\Phi >_{\min} \psi)(\omega^*) = 1$ iff $V^*(\psi)(\omega^{*\prime}) = 1$ for the least $\omega^{*\prime} \in \omega^*{\uparrow}\Phi$
c. $V^*(\Phi >_\exists \psi)(\omega^*) = 1$ iff $V^*(\psi)(\omega^{*\prime}) = 1$ for some $\omega^{*\prime} \in \omega^*{\uparrow}\Phi$

By referring to the "least" $\omega^{*\prime}$ in the set I assume the obvious order of the sequences, i.e., in terms of the position in $\omega^*$ at which they start. It can then be shown that (6a) yields the desired prediction for the probabilities of conditionals with interrogative antecedents: $P(\Phi > \psi) = P(\bigwedge_{\varphi\in\Phi}(\varphi > \psi))$.

Conditionals with disjunctive antecedents are also typically interpreted by distribution over the disjuncts – that is, the probability of ($\varphi_1$ or $\varphi_2$) $> \psi$ is the probability of the conjunction $(\varphi_1 > \psi) \wedge (\varphi_2 > \psi)$. This likewise falls out if we assume that the disjunctive antecedent denotes a set of propositions (a commonly made assumption in Inquisitive Semantics, cf. [4]) and interpreted according to (6a). In contrast, (6b) corresponds to an interpretation that gives the disjunction in the antecedent its Boolean interpretation. Finally, (6c) also distributes over the elements of $\Phi$, but yields the probability of the disjunction, rather than the conjunction, of the conditionals.

For the simple special case in which the antecedent has two alternatives, neither of which contains conditionals, the resulting readings are as follows (the general case is similar but more tedious to show):

(7) a. $P(\{A, B\} >_\forall C) = P((A > C) \wedge (B > C))$
b. $P(\{A, B\} >_{\min} C) = P((A \vee B) > C)$
c. $P(\{A, B\} >_\exists C) = P((A > C) \vee (B > C))$

For reasons of space I cannot provide detailed proofs here. Suffice it to point out that (7a) is the probability that the first $A$-world is a $C$-world *and* the first $B$-world is a $C$-world; (7b) is the probability that the first world at which either $A$ or $B$ is true is a $C$-world; and (7c) is true iff the first $A$-world is a $C$-world *or* the first $B$-world is a $C$-world (or both). Clearly (7a-c) asymmetrically entail each other. Whether all three are attested as potential readings for conditionals with interrogative or disjunctive antecedents is an open empirical question.[4]

## References

1. Adams E (1965) The logic of conditionals. Inquiry 8:166–197
2. Adams E (1975) The logic of conditionals. Reidel
3. Bennett J (2003) A philosophical guide to conditionals. Oxford University Press
4. Ciardelli I, Roelofsen F (2009) Generalized inquisitive semantics and logic. http://sites.google.com/site/inquisitivesemantics/. Accessed Nov 2009
5. de Finetti B (1935) La logique de la probabilité. In: et Cie H (ed) Actes du Congrès International de Philosophie Scientifique, Paris, Paris, pp IV 1–IV 9

[4] I am grateful to two anonymous reviewers for helpful feedback.

6. Edgington D (1995) On conditionals. Mind 104(414):235–329
7. Eells E, Skyrms B (eds) (1994) Probabilities and conditionals: belief revision and rational decision. Cambridge University Press
8. Gilio A, Sanfilippo G (2013) Conjunction, disjunction and iterated conditioning of conditional events. Adv Intell Syst Comput 190:399–407
9. Gilio A, Sanfilippo G (2014) Conditional random quantities and compounds of conditionals. Studia Logica 102:709–729
10. Hájek A, Hall N (1994) The hypothesis of the conditional construal of conditional probability. In: [6], pp 75–110
11. Hamblin CL (1958) Questions. Australas J Philos 36(3):159–168
12. Hamblin CL (1973) Questions in montague English. Found Lang 10:41–53
13. Jeffrey RC (1991) Matter-of-fact conditionals. In: The symposia read at the joint session of the aristotelian society and the mind association at the university of Durham, The Aristotelian Society, pp 161–183, supplementary Volume 65
14. Karttunen L (1977) Syntax and semantics of questions. Linguist Philos 1:3–44
15. Kaufmann S (2004) Conditioning against the grain: abduction and indicative conditionals. J Philos Logic 33(6):583–606
16. Kaufmann S (2005) Conditional predictions: a probabilistic account. Linguist Philos 28(2):181–231
17. Kaufmann S (2009) Conditionals right and left: probabilities for the whole family. J Philos Logic 38:1–53
18. Kaufmann S (2015) Conditionals, conditional probability, and conditionalization. In: Zeevat H (ed) Schmitz HC. Bayesian natural language semantics and pragmatics. Springer, pp 71–94
19. Khoo J (2016) Probabilities of conditionals in context. Linguist Philos 39:1–43
20. Kratzer A (1981) The notional category of modality. In: Eikmeyer HJ, Riesner H (eds) Words, worlds, and contexts, Walter de Gruyter, pp 38–74
21. Kratzer A (1991) Conditionals. In: von Stechow A, Wunderlich D (eds) Semantik: Ein internationales Handbuch der zeitgenössischen Forschung. [=Semantics], de Gruyter, pp 651–656
22. Kratzer A (2012) Modality and conditionals. Oxford University Press
23. Lance M (1991) Probabilistic dependence among conditionals. Philos Rev 100:269–276
24. Lewis D (1976) Probabilities of conditionals and conditional probabilities. Philos Rev 85:297–315
25. Lewis D (1986) Postscript to “Probabilities of conditionals and conditional probabilities”. In: Philosophical Papers, vol 2, Oxford University Press, pp 152–156
26. McGee V (1989) Conditional probabilities and compounds of conditionals. Philos Rev 98(4):485–541
27. Mellor DH (ed) (1990) Philosophical papers: F. Cambridge University Press, P. Ramsey
28. Oaksford M, Chater N (2003) Conditional probability and the cognitive science of conditional reasoning. Mind Lang 18(4):359–379
29. Oaksford M, Chater N (2007) Bayesian rationality: the probabilistic approach to human reasoning. Oxford University Press, Oxford, UK
30. Over DE, Evans JSBT (2003) The probability of conditionals: the psychological evidence. Mind Lang 18(4):340–358
31. Ramsey FP (1929) General propositions and causality. Printed in [28], pp. 145–163
32. Stalnaker R, Jeffrey R (1994) Conditionals as random variables. In: [6], pp 31–46
33. van Fraassen BC (1976) Probabilities of conditionals. In: Harper WL, Stalnaker R, Pearce G (eds) Foundations of probability theory, statistical inference, and statistical theories of science, The University of Western Ontario Series in Philosophy of Science, vol 1, D. Reidel, pp 261–308

# Dynamic Analysis of the Development of Scientific Communities in the Field of Soft Computing

Ekaterina Kutynina and Alexander Lepskiy

**Abstract** This paper is dedicated to the research of the dynamics of development and interactions among several scientific communities in the field of fuzzy logic and soft computing. This analysis was performed with the help of the following characteristics: conferences participants' renewal, the level of cooperation in scientific communities, participation of one community's key players in activities of the other ones, comparative number of most active participants in each community, uniformity of key players' participation in different conferences.

**Keywords** Scientific communities · Key participants of communities · Interaction between scientific communities

## 1 Introduction

At present scientific communities are an essential part and an important form of the scientific process organizing. In recent years, scientific communities are often studied by methods of network analysis. In particular, the co-authorship networks and citing networks [7] are popular. However, the scientific community tends to develop: there are new communities; some communities degrade, while others are combined, etc. The life cycle of scientific communities is considered in a number of works (see [1]). The interactions among the scientific communities in the field of artificial intelligence for the last 19 years were investigated in [2, 4] is a similar study that was carried out for scientific communities in the field of computer science. In [10] the dynamic changes in the co-authorship network of conference ISIPTA [9] were analyzed.

E. Kutynina · A. Lepskiy (✉)
Higher School of Economics, 20 Myasnitskaya Ulitsa, Moscow 101000, Russia
e-mail: alex.lepskiy@gmail.com

E. Kutynina
e-mail: ekytinina@gmail.com

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_37

The given work is devoted to the investigation of the development and interactions of scientific communities in the field of fuzzy mathematics (EUSFLAT, NAFIPS), imprecise probability (SIPTA, BFAS) and soft computing (SMPS) during the period 1999–2014. The database of this study is based on the materials of the conferences held by the above mentioned scientific communities.

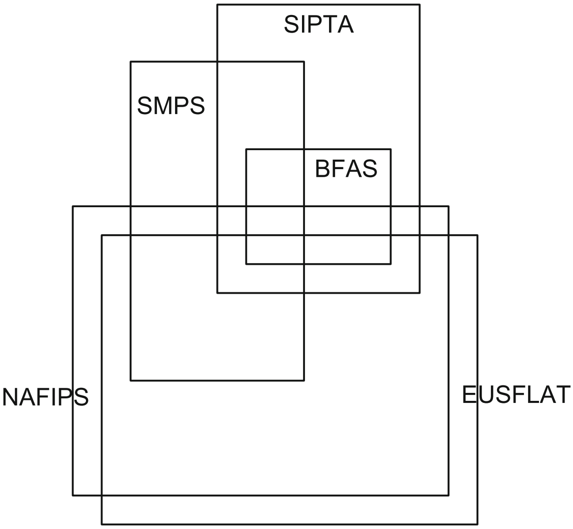
## 2 Dataset Description

Following scientific communities were considered:

- BFAS (Belief Functions and Applications Society) [3]. BFAS was formed in 2010. Conference—BELIEF.
- EUFSLAT (European Society for Fuzzy Logic and Technology) [8] was founded in 1998. Conference—EUFSLAT.
- NAFIPS (North American Fuzzy Information Processing Society) [6]. NAFIPS was established in 1981. Conference—NAFIPS.
- SIPTA (The Society for Imprecise Probability: Theories and Applications) [9] was formed in 2002. Conference—ISIPTA (International Symposium on Imprecise Probability: Theories and Application).
- SMPS (International Conferences on Soft Methods in Probability and Statistics) [5]. Conference SMPS has been held since 2002.

Conferences EUFSLAT, ISPITA, SMPS, BELIEF (for brevity they are designated with letters E, I, S, B respectively) are held once every 2 years, and conference NAFIPS (symbol N)—every year.

**Fig. 1** The schematic visualization of the intersection of communities' themes



The Fig. 1 provides the visualization of the intersection of the conferences' themes. It could be expected, that connection within groups of communities EUSFLAT, NAFIPS, SMPS on the one hand, and BFAS, ISIPTA on the other hand would be tighter within groups than between them.

The data about the authors of the papers presented at conferences during the period 1999–2014 was collected. There are $N = 3377$ participants in total. Since almost all the conferences are held every 2 years, the entire time interval 1999–2014 was divided into 8 equal subintervals [1999, 2001), …, [2013, 2015). On the figures below the left boundaries of all subintervals are indicated on the horizontal axes. Data on the conferences held by the NAFIPS community have been combined for the 2009 and the 2010, 2011 and 2012 and then considered as a single event.

# 3 Analysis of the Development Dynamic of Scientific Communities

## 3.1 Renewal of Conferences' Participants

One of the main indicators that can characterize the internal development of the scientific community can be a number of new conference participants, who did not participate in previous conferences of the community. Let $All_i^j$ is a set of all participants of the conference $j$ in the period $i$. The coefficient of renewability for $j$th conference in the period $i$ can be considered as a value $U_i^j = \frac{\left|All_i^j \backslash \left(All_1^j \cup \ldots \cup All_{i-1}^j\right)\right|}{\left|All_i^j\right|}$, which is the ratio of the number of new entrants to the number of all participants of this conference in the considered period.

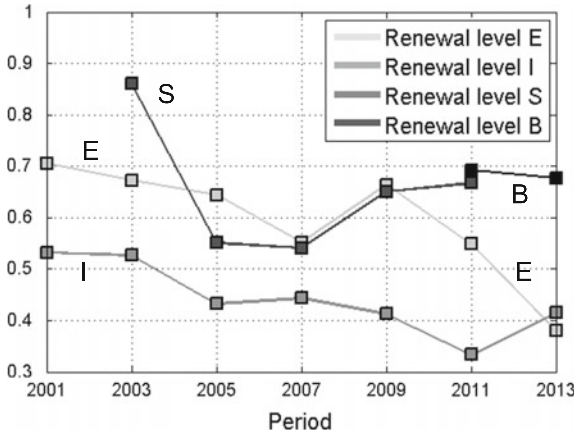


**Fig. 2** The renewal level of participation in conferences

The Fig. 2 shows that the average renewal of ISIPTA conference participants is significantly lower than renewability for other conferences. Almost all the conferences (except for SMPS) tend to decrease renewal of participants. At the same time it should be noticed that the total number of participants in each conference on average varies slightly.

Experienced researchers take part in the conferences as well as their young colleagues. However, in terms of development and interaction of the scientific communities it seems more meaningful to consider the information about experienced researchers, in other words those who took part in several conferences and presented several papers at the same conference. Let's call these researchers key participants.

### 3.2 Key Participants

The significance of a participant s is defined as $Val_i(s)$, the sum of the researchers contributions in the creation of all publications for the period $i$, where $s = 1, \ldots, 3377, i = 1, \ldots, 8$. It will be assumed that the total value of the publication is equal to 1 and is divided among all co-authors equally. If the participant $s$ took part in several conferences during the period $i$ his total contribution is calculated as $Val_i(s) = \sum_j Val_i^j(s)$ (the total contribution was calculated separately for two periods 1999–2007 and 2009–2015, since the number of conferences which were held during this periods was different). Those conference participants are called key participants, whose total contributions exceed a certain limit $p$. Below are the results for the cut-off threshold of the key participants $p = 2$, in other words, those who totally wrote not less than 2 works over 8 years. The set of key participants in the period $i$ is indicated $K_i$.

Suppose that $K_i^j = K_i \cap All_i^j$ is a set of key participants of the community $j$, which held the conference in the period $i$. In this case there is an opportunity to study the dynamics of the $K_i^j$ sets structural changes and characteristics of their interactions. The Fig. 3 shows the changes in the total numbers of the key participants in all communities $K_i$. As it can be seen from the graph, the rate tends to increase in the considered time interval. This suggests that interest in research in the field of imprecise probability, fuzzy sets and soft computing increases with time, the significance of this research area is growing.

### 3.3 The Level of Communication Among Communities in Relation to the Key Participants

One of the key issues is to determine the level of cooperation in scientific communities in connection with the common key participants of the conferences organized by thematically close communities. This level can be defined as the correlation between

**Fig. 3** Dynamics of changes in the total number of communities' key participants

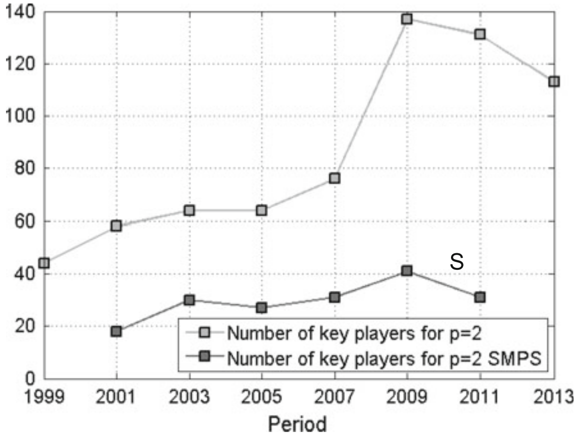
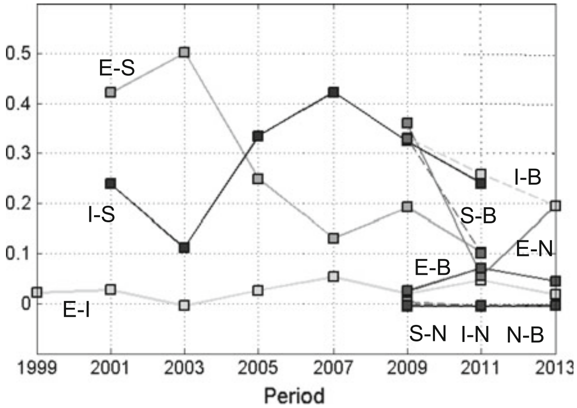
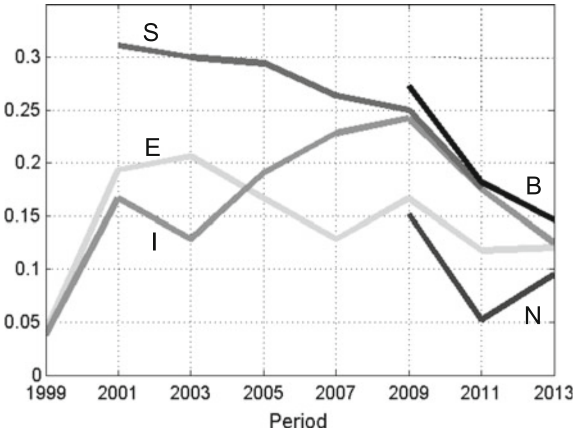


**Fig. 4** The correlation between the key participants of the conferences



the vectors of the significances of the key participants in couples of communities. Thus, $n_i = |K_i|$ is the number of the key participants in the period $i$. For each period $i$, the vector $\mathbf{w}_j^i = (w_{1j}^i, \ldots, w_{n_i j}^i)$ is put in correspondence to conference $j$, where $w_{sj}^i = Val_i^j(s)$ is the significance of scientist $s$ just for the conference of $j$th community in the period $i$. Then, the level of cooperation among communities $k$ and $j$ in the period $i$ can be considered as a selective linear Pearson correlation coefficient $r_{kj}^i$ between vectors $\mathbf{w}_k^i$ and $\mathbf{w}_j^i$. The Fig. 4 shows the variation of the selective correlation coefficient for all pairs of communities. It is evident that, as a rule, the level of cooperation among communities in terms of the common key participants of conferences either are initially small (for example, among E and I, E and B), or have a tendency to a decrease (for example, between E and S, S and B, I and B). All this suggests a trend to isolate these communities. The exception here is a pair of I and S.

**Fig. 5** The dynamics of changes in the participation rate in other communities



### *3.4 The Participation of the Key Participants in Other Communities*

If, however, the key participants of the community are involved in other conferences, the extent of such participation for the community $j$ in the period $i$ can be estimated with the formula $k_j^i = \frac{1}{l \cdot n_j^i} \sum_{k=1,k\neq j}^{5} m_{jk}^i$, where $l$ is the number of non-empty sets $K_i^j$ in the $i$th period, $m_{jk}^i = \left| K_i^j \cap K_i^k \right|$ is the number of common key members of communities $j$ and $k$ in the period $i$, $n_j^i = \left| K_i^j \right|$ is the number of key conference participants of $j$th community in the period $i$. The higher this ratio, the more actively the key participants of the particular community are involved in other communities' activities. The high value of this factor could mean that the key participants do not regard the community as a key community in the considered field of knowledge.

The Fig. 5 is a visualization of this ratio dynamics. One can see that the least "key" one was the community of S until 2009. The community N turned to be the most "closed", in other words, the key members of this community rarely visit other conferences. But this can be explained by "regional" separateness of this community. The most stable is the community E, for which the rate of participation of key scientists in the other communities does not change much and remains quite small.

### *3.5 The Most Active Community Members and Most Active Communities*

Suppose $K_i$ is a set of the key participants of all conferences for the period $i$, $K = \bigcup_i K_i$ is a set of the key participants of all conferences on record.

Consider the “friendship” graph for the communities participants (conferences) $G_i = (K_i, E_i), i = 1, \ldots, 8$ and $G = (K, E)$, where $E_i$ $(E)$ is the set of edges with weights $e_{st}$, which is equal to the number of the joint participation of key participants $s$ and $t$ in the same conferences for the period $i$ (for all periods).

In this connection we can rise the problem of determining those members who are “friends” with the greatest number of other key participants, taking into account not only direct relations (participation in one conference), but also indirect (i.e. a “sign through a friend”). Such participants can be considered as the most active members of the communities. This problem is solved in the analysis of network structures with the help of the eigenvector centrality. The calculation of the measure of centrality for each node is connected with the solution of eigenvalue finding problem regarding the adjacency matrix $A$ of the network graph [7]: the vector of the relative centralities $\mathbf{x}$ is an eigenvector of the adjacency matrix that corresponds to the largest eigenvalue $\lambda_{\max}$.

On the basis of the eigenvector centrality indicators such as the average value of activities of key participants in each community were introduced: $act_j = \frac{1}{m_j N_j} \sum_{s=1}^{N} n_{sj} x_s$, where $x_s$ is $s$th component of the relative centralities vector $\mathbf{x} = (x_1, \ldots, x_N)$ of the “friendship” graph of key community members; $n_{sj}$ is the number of times that the participant $s$ took part in the conference $j$, $j = 1, \ldots, 5$; $N_j$ is the total number of key participants of the community $j$ at the moment of the index calculating; $m_j$ is the number of conferences that had been held by the community $j$ by the moment of index calculating.

The Table 1 provides the list of most active members of all communities. The number of the considered scientists’ participation in the conferences of each community is shown in brackets in the last column.

Dynamics of changes in the average value of activity for all communities is represented in Fig. 6. The most active key players are participants of SMPS community, the lowest average activity is observed among the participants of NAFIPS and BFAS communities.

**Table 1** The list of the most active participants of all communities

| Key participant | Centrality | Participation in communities |
|---|---|---|
| Dubois D. | 0.260 | E(7), I(5), S(5), B(2) |
| Kacprzyk J. | 0.256 | E(8), S(5), N(2) |
| Grzegorzewski P. | 0.229 | E(6), S(6), N(1) |
| de Baets B. | 0.211 | E(8), I(1), S(4) |
| Trillas E. | 0.183 | E(7), N(3) |
| Prade H. | 0.181 | E(6), I(1), S(3) |
| Novák V. | 0.175 | E(8), N(2) |
| Recasens J. | 0.172 | E(7), S(1), N(2) |

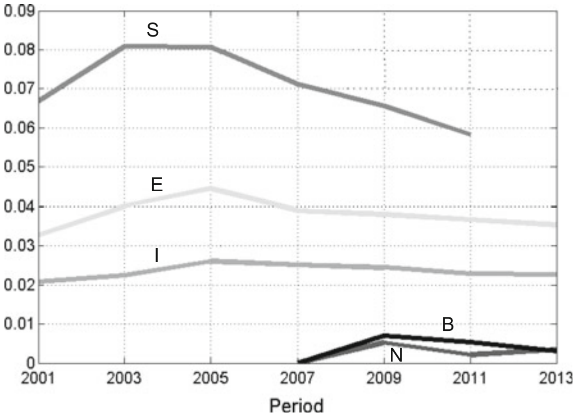


**Fig. 6** The dynamics of changes in the average value of activity for the all communities

## 3.6 Analysis of Participation Uniformity of the Key Participants in Different Communities

Each participant $s$ was assigned with the vector $\mathbf{n}_s = (n_{s1}, \ldots, n_{s5})$, where $n_{sj}$ is the number of times, which the participant $s$ took part in the conference $j$, $j = 1, \ldots, 5$. The vector $\mathbf{n}_s = (n_{s1}, \ldots, n_{s5})$ was put to correspondence to the vector of relative frequencies $\mathbf{p}_s = (p_{s1}, \ldots, p_{s5})$, where $p_{sj} = \frac{n_{sj}}{\sum_{k=1}^{5} n_{sk}}$. Then $\mathbf{p}_s = (p_{s1}, \ldots, p_{s5})$ is some probability distribution. Pose the question of the non-uniformity degree of this distribution, which characterizes the degree of heterogeneity of participation in conferences of different communities for the $s$th scientist. This degree can be estimated with the help of the Shannon entropy $S(\mathbf{p}_s) = -\sum_{j=1}^{5} p_{sj} \log_2 p_{sj}$. For the uniform distribution this function reaches its maximum $S\left(\frac{1}{5}, \ldots, \frac{1}{5}\right) = \log_2 5$. The entropy achieves the minimum $S(\mathbf{p}) = 0$ when exactly one of the $p_j$ is one and all the rest are zero.

Now, for each set of key participants $K^j$ of community $j$ the average homogeneity was calculated by the formula $unif_j = \frac{1}{|K^j|} \sum_{s \in K^j} S(\mathbf{p}_s)$.

Great value of $unif_j$ indicates that the key members of this community are also actively involved in the work of other communities. A small value of $unif_j$ demonstrates a certain “isolation” degree of the community. The Fig. 7 is a graphical representation of communities by points on the plane, where the first coordinate on the horizontal axis is the average homogeneity of the community, and the second one on the vertical axis is the average value $\overline{Val^j}$ of the aggregate contributions of the key participants in community $j$ for the entire considered period. It can be seen, that the most “closed” communities are EUSFLAT and ISIPTA, thus the average contribution of key participants of these conferences is the highest. The most open community is SMPS, which can be explained by the variety of scientific papers themes presented

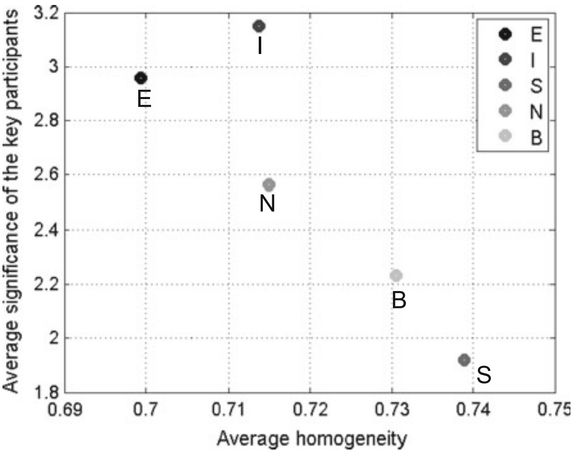


**Fig. 7** The average homogeneity and average significance of the key communities' participants

at conferences of this community. On the other hand it is evident that these two characteristics—average contribution and uniformity are strongly correlated. Again, the "outlier" here is ISIPTA community.

## 4 Conclusions

The main conclusions of this research are as follows:

- almost all of the conferences (except for SMPS) have a tendency to reduce the renewal of its members (at a fairly constant total number of conference participants); on average the renewal of ISIPTA conference participants is significantly lower than other conferences renewal;
- the level of cooperation in scientific communities in relation to the common key participants of the conferences either is initially small (for example, between EUSFLAT and ISIPTA, EUSFLAT and the BELIEF), or have a tendency to a decrease (for example, between EUSFLAT and SMPS, SMPS and BELIEF, ISIPTA and BELIEF). All this says about the trend to isolate these communities; exceptions here are ISIPTA and SMPS conferences;
- in terms of the participation of the key participants of a particular community in the activities of other communities, until 2009 the most "open" was a conference SMPS; as far as this characteristics is concerned the most stable community is EUSFLAT, for which the participation rate of key scientists in other communities does not change much and remains quite small;
- the most active participants of the communities were emphasized; It shows that the most active participants are key participants of SMPS community; the lowest activity was observed among the participants of NAFIPS and BFAS communities;

- in terms of uniformity of participation of key participants of a particular community in other communities, the most “closed” communities are EUSFLAT and ISIPTA, thus the average contribution of key participants of these conferences is the highest; the most open community on this indicator is SMPS.

**Acknowledgments** The financial support from the Government of the Russian Federation within the framework of the implementation of the 5-100 Programme Roadmap of the National Research University Higher School of Economics is acknowledged.

## References

1. Belák V, Karnstedt M, Hayes C (2011) Life-cycles and mutual effects of scientific communities. Procedia—Procedia Soc Behav Sci 00:36–47
2. Belák V, Hayes C (2015) The risks of introspection: a quantitative analysis of influence between scientific communities. In: Proceedings of the 28th international florida artificial intelligence research society conference, pp 8–13
3. Belief functions and applications society. http://www.bfasociety.org/
4. Biryukov M, Dong C (2010) Analysis of computer science communities based on DBLP. In: Research and advance technologies for digital library. Springer, Berlin, pp 228–235
5. Conference on soft methods in probability and statistics. http://smps2014.ibspan.waw.pl
6. North American fuzzy information processing society. http://nafips.ece.ualberta.ca
7. Newman MEJ (2010) Networks: an introduction. Oxford University Press, Oxford
8. The European society for fuzzy logic and technology. http://www.eusflat.org/
9. The society for imprecise probability: theories and applications. http://www.sipta.org
10. Walter G, Jansen C, Augustin T (2015) Updated network analysis of the imprecise probability community based on ISIPTA electronic proceedings. In: Proceedings of the 9th ISIPTA, Pescara Italy, p 351

# Talk to Your Neighbour: A Belief Propagation Approach to Data Fusion

**Eleonora Laurenza**

**Abstract** Data fusion is a major task in data management. Frequently, different sources store data about the same real-world entities, however with conflicts in the values of their features. Data fusion aims at solving those conflicts in order to obtain a unique global view over those sources. Some solutions to the problem have been proposed in the database literature, yet they have a number of limitations for real cases: for example they leave too many alternatives to users or produce biased results. This paper proposes a novel algorithm for data fusion actually addressing conflict resolution in databases and overcoming some existing limitations.

**Keywords** Data fusion · Bayesian networks · Belief propagation · Data uncertainty · Data integration

## 1 Context and Motivation

Data fusion is the task of merging multiple representations of the same real-world entities in order to obtain a single and unified view of them. In relational database systems, data are represented by records (tuples) in tables and are characterized by a multiplicity of features. Some features are referred to as key, since they uniquely identify the records. One major problem of the various representations of the same entity in different data sources happen to have disagreeing values for corresponding features, data fusion involves detecting and solving such conflicts [2].

The problem has an increasingly significant industrial relevance, because of the massive proliferation of redundant and often contradictory data. Moreover, the complexity of the most recent data management scenarios (statistical microdata, genomic data, linked open data), together with the always increasing volumes, cause quality loss and reduced trust in the data [15]. Database fusion problem aims at achieving a unified view of various representations of the same entity by solving the conflicts among the disagreeing features. In order to fuse databases, other activities are needed,

E. Laurenza (✉)
Sapienza University, Piazza Aldo Moro, 5, Rome, Italy
e-mail: eleonora.laurenza@uniroma1.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_38

which in the literature are typically grouped in the *data integration* problem [18]. It involves *schema integration* [1, 9] and *data matching* [5]. The former aims at fusing the databases at a schema level, hence achieving the same logical representation of entities, that is, the same name for relations and features; the latter concerns the identification of the same real-world entities in the different sources, as it is often the case that common identifiers (such as social security numbers for individuals, VAT code for companies) are not present. Data fusion is a meaningful problem in database literature, however the results that have been provided have proven to be not effective in many real cases. Several algorithms simply ignore the conflicts (*conflict-ignoring*), leaving the choice to the final users; other approaches adopt a preference strategy (*conflict-avoiding*), taking the value from the most trustworthy sources. Finally, some others actually try to solve the conflicts (*conflict-solving*), but with techniques that are limited to simple arithmetic approximations [3]. These approaches have a number of limitations. Ignoring or avoiding conflicts is not practical, especially with the recent explosion of available sources and attributes for each entity. Users would be exposed to a very large number of alternatives for each conflict. Algorithms based on approximations only lead to local bests, since the specific kind of approximation depends on each user's sensitivity, overall resulting in a biased global view.

This work proposes *BP-fuse* (*Belief Propagation fusion*), a novel algorithm for solving conflicts in database fusion. This is the first approach that uses the probabilistic dependencies among attributes exploiting the non-conflicting values to choose the "true" values for the conflicting ones. The probabilistic dependencies are modeled using Bayesian networks for a compact representation and efficient querying.

## 2 The Approach

Let us approach the problem of data fusion by referring to the real application of several European company registers, which are collections of records about multinational enterprises in EU, considering two of them, held, for example, by two different national statistical institutions of the respective member states: Italy and Germany. The registers are modeled as two tables. Figure 1 shows a fragment of those tables.

**Fig. 1** Sample tables from European business registers

ITALIAN BUSINESS REGISTER

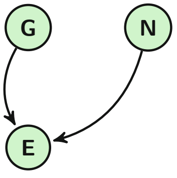
| ID | L_NAME | EMP_NO | GEO_AREA | NACE | PROFIT |
|---|---|---|---|---|---|
| 526 | FCA | 100k | Ur | AUTO | 20M |
| 114 | SIEMENS | 360k | Co | ICT | 700M |
| 834 | Ferrari | 9k | - | AGRI | 200M |

GERMAN BUSINESS REGISTER

| ID | L_NAME | REV | GEO_AREA | EMP_NO | FORM |
|---|---|---|---|---|---|
| 38 | FCA | - | Ur | 200 | SPA |
| 73 | SIEMENS | 6.14G | Ur | 100 | Gmbh |
| 714 | LVMH | 3.06G | Co | 83k | - |

For one single company some characteristics are known in the Italian register and unknown in the German one and viceversa. Besides, for two companies, the two registers have conflicting values for the corresponding attributes. The goal is obtaining a unified business register by fusing the two. For each of the registers, ID is the primary key and L_NAME is the legal identifier of the company. Both the registers store the geographical area (GEO_AREA) and the number of employees (EMP_NO). My approach relies on Bayesian networks to solve the conflicts. They are DAGs (Directed Acyclic Graphs) that specify a multivariate joint probability distribution over a set of random variables used to represent knowledge in an uncertain domain [13]. The *nodes* represent the random variables that are concerned in the reality of interest. Probabilistic dependencies among variables are graphically expressed by *directed edges* in the network. Each node is labelled with a *conditional probability distribution* (CPT) table. It contains the distribution of such variable, as it is conditioned on all the variables corresponding to incoming edges and encodes the quantitative knowledge about the domain. In Bayesian networks, we need to specify the graph topology and the values of each CPT. It is possible to infer both of these automatically from data (for example there are simple strategies to learn CPT's from training sets using the ML or EM algorithms) or exploiting the knowledge of the domain experts [12].

For the domain in the example, a simple net is shown in Fig. 2: the net represents some kind of causal dependency relating *G* and *N* with *E* and the CPT refers to node E and shows how the number of employees varies depending on the geographical area and the economic classification of the company. The geographical area where the production site of the company resides, together with the economic classification of its business are reported to influence the number of employees as shown in the probability table in Fig. 2. For instance, automotive enterprises (AUTO) situated in the country (Co) tend to have between 10 and 49 employees, while construction enterprises (CONST) in urban centers (Ur) have about 70 employees with a probability of 0.33. Let us consider the fusion of the two records referring to the FCA in Fig. 1. FCA is present in both the registers, the attributes NACE and PROFIT are present only in the Italian register: therefore values AUTO and 20M are directly in the result. REV and FORM, which are present only in the German register, appear with their values in the result as well (Fig. 3).



| | **GEO by NACE by EMP_NO** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AGRI | | AUTO | | CONST | | ICT | |
| | Ur | Co | Ur | Co | Ur | Co | Ur | Co |
| $< 10$ | 0.6 | 0.01 | 0.2 | 0.23 | 0.25 | 0.19 | 0.01 | 0.75 |
| 10-49 | 0.34 | 0.1 | 0.03 | 0.4 | 0.3 | 0.34 | 0.01 | 0.21 |
| 50-249 | 0.03 | 0.32 | 0.12 | 0.07 | 0.33 | 0.43 | 0.2 | 0.03 |
| $> 249$ | 0.03 | 0.57 | 0.65 | 0.3 | 0.12 | 0.04 | 0.78 | 0.01 |

**Fig. 2** Relations among GEO, NACE and EMP_NO

| EUROPEAN BUSINESS REGISTER | | | | | | | |
|---|---|---|---|---|---|---|---|
| ID | L_NAME | EMP_NO | GEO_AREA | NACE | REV | PROFIT | FORM |
| 1 | FCA | 100k | Ur | AUTO | - | 20M | SPA |
| 2 | SIEMENS | 360k | Ur | ICT | 6.14G | 700M | GMBH |
| 3 | Ferrari | 9k | Co | AGRI | - | 200M | - |
| 4 | LVMH | 83k | Co | - | 3.06G | - | - |

**Fig. 3** The result of BP-fuse algorithm

The two relations agree on the GEO_AREA, but show a conflict for EMP_NO: 100k for the Italian one, 200 for the German one. BP-fuse solves conflicts of this kind, by evaluating the plausibility of the candidate values, given the certain ones. Using the simple Bayesian net in Fig. 2 with only three variables, the algorithm calculates $P(100k \mid \text{Ur, AUTO})$, which is 0.65; it also calculates $P(200 \mid \text{Ur, AUTO})$, yielding 0.12. The most plausible value is 100k and it is assigned to EMP_NO in the fused record. The case for SIEMENS is quite similar, however particular attention must be paid as both GEO_AREA and EMP_NO disagree. The final two records, Ferrari and LVMH, appear only in one relation and so they are directly part of the result. Figure 4 shows a more complete network for this example, including variables (J_LABOUR_COST, EXPORT_VOL) that are not attributes of the input tables, but are meaningful in the domain of interest. The progress bars in each node intuitively represent the marginal probabilities for each value of the random variables, when all the dependencies are considered and after the network convergence.

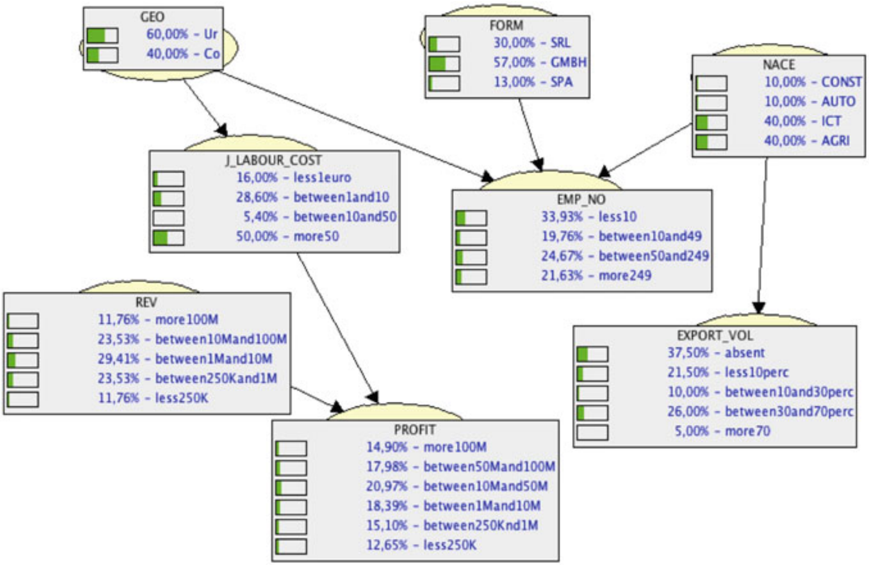


**Fig. 4** A larger Bayesian network

# 3 BP-fuse

Simple sensor model (SSM) is the conceptual data model envisaged to support data fusion in this paper and gives a solution independent of the relational model, even though the correspondence between the constructs of relational model and those of SSM is quite straightforward. In this model, data to be fused are modeled as the measures in a physical sensor. A *sensor* $S(I, \mathbf{V})$ is characterized by an identifier $I$ and a set of variables $V = V_1, \ldots, V_n$. The identifier and the variables represent the attributes of the entity measured by the sensor, in particular the identifier is the real-world name. The instances of each sensor are named *measures*, where each one is an assignment $i$ for $I$ and $v = (v_1, \ldots, v_n)$ for $\mathbf{V}$. SSM also comprises the information about the causal dependencies among the variables of the sensors and adopts constructs from Bayesian networks to model them. The identifiers are also a link between different sensors, because they allow to tell what measures refer to the same entity. Given two sensors $S_1(I, V_1)$ and $S_2(I, V_2)$, where $V_1$ and $V_2$ are two sets of variables, their fusion is a third sensor $S_3(I, V_3)$, such that for each pair of matched measures (they have the same value for the identifier) showing a conflict on a common variable, the conflict is solved in $S_3$. In particular, one single value for that variable is chosen from one of the two sensors. In relational terms, this corresponds to a JOIN between $S_1$ and $S_2$ on the common variable $I$, with the application of some conflict-solving function on the variables, and is summarized by the following SQL query:

```
SELECT S1.I, fuse(S1.V1, S2.V1), ..., fuse(S1.Vn, S2.Vn)
FROM S1,S2  WHERE S1.I = S2.I
```

Here `fuse()` denotes a conflict solving function, using, for example, BP-fuse. This approach has two phases: the former, *emission*, is devoted to the extraction of the measures from the input sensors; the latter, *unification*, has the responsibility to actually solve conflicts among the values of the variables in all the sensors. For every sensor and for every measure $m(i, v_1, \ldots, v_k)$, the emission phase produces a set of triples $(i, V_1, v_1)$, ..., $(i, V_k, v_k)$. The triples are then grouped by identifier $i$ into *candidate entities* (CE), which are collections of triples referring to the same real-world entity. In a candidate entity the triples are in turn grouped by $V_i$ into *candidate sets* (CS). A candidate set collects for each variable and entity, all the possible values coming from different measures and sensors.[1] The unification phase has the responsibility to produce from every candidate entity a measure for $S_r$. To achieve this, BP-fuse needs to reduce every candidate set to a unique value.

Four cases are possible with respect to candidate set reduction: (i) *there is only one non-null value in the candidate set*: BP-fuse chooses the non-null candidate value; (ii) *null set*: the candidate set only contains the null value, BP-fuse chooses the null value; (iii) *no conflict*: the candidate set has exactly one value, BP-fuse chooses

[1] Notice that for a given $i$, different candidates for a variable can also derive from the same sensor, in case of duplicate measures.

this value; (iv) *conflict*: there are different values in the candidate set. Case (iv) is indeed very common and, moreover, several variables are likely to be conflicting in a measure at the same time.

For every candidate entity, BP-fuse considers all the conflicting variables. Let $V_1, \ldots, V_t$, be such variables. BP-fuse generates all the possible assignments $a = (v_1, \ldots, v_t)$, where $v_i$ is chosen from candidate set $V_i$. Then the algorithm investigates the plausibility of each assignment $a$ as follows. Let $V_{t+1}, \ldots, V_q$ be the other variables of the measure, the ones for which the respective candidate sets have already been reduced by applying cases i-iii. For each assignment $a$, BP-fuse estimates the plausibility with the support of the associated Bayesian net. It generates and evaluates queries such as:

$$P(v_1, \ldots, v_t \mid v_{t+1}, \ldots, v_q) = \frac{P(v_1, \ldots, v_t, v_{t+1}, \ldots, v_q)}{P(v_{t+1}, \ldots, v_q)} \tag{1}$$

In order to efficiently compute the lhs of (1) for $a$, BP-fuse applies some basic manipulations, resulting in the rhs. Each conjunctive form in the rhs is factorized into $P(v_1)P(v_2 \mid v_1) \ldots P(v_n \mid v_{n-1}, \ldots, v_1)$ by applying the chain rule. Now, BP-fuse orderly calculates each factor $P(v_i \mid v_1, \ldots, v_j)$ by applying an algorithm for the network convergence such as *belief propagation* [13]. It starts from initial factors $P(v_i)$ of the chain and then uses each $v_i$ in the evidence set for the following factors. It eventually extracts the marginal probability $(V_i = v_i)$ after the network convergence for the conditioned variable $v_i$. BP-fuse calculates the plausibility of $a$ by replacing previously calculated factors in (1). At this step, BP-fuse chooses for the candidate entity under consideration the assignment $a$ with top plausibility. It reduces every candidate set to a unique value and, as a consequence, produces a measure for $S_r$. The application of the explained steps to all the candidate entities results in the generation of all the fused measures for $S_r$. BP-fuse returns the measure corresponding to the assignment with the highest plausibility, solving the conflicts together. One recognized way to evaluate algorithms for data integration and data fusion in particular, consists in weighing the data fusion answer by means of two indicators: *completeness* and *conciseness* [3] with an approach recalling the more usual terms of precision and recall.

A good fusion algorithm would be expected to increase the completeness and, at least, not to decrease the conciseness with duplicates. For the running example, the result has the best values for intentional completeness as it contains the union of all the variables. BP-fuse maximizes extensional completeness as well, since the key-value pairs are generated for all the involved sensors and no measures are discarded during the unification phase. BP-fuse maximizes also the intensionally conciseness by allowing for the application of any schema matching algorithm. Regarding extensional conciseness, the algorithm also gives the highest value: the emission produces a key-value pair for each measure and variable, and the unification phase collects all the pairs with the same real-world key into a single fused measure.

## 4 Related Work

In the database literature some techniques for fusion have been provided. Some solutions rely on relational algebra operations [7, 14, 21], unaffordable in many real cases. Others actually try to solve the conflicts and propose a combination of the disagreeing values based on simple arithmetics or user-defined functions [2, 3]. Their results are not always acceptable, as, for instance, the average of two conflicting values may be out of the acceptable domain or, in any case, tightly coupled to each user's sensitivity.

In multi sensor fusion, indeed, the problem consists in combining sensor data deriving from disparate sources, with uncertain information about the specific scenario, which can be any, including: combining two-dimensional images from two cameras at slightly different viewpoints, combining animal tracking data with meteorological and animal habitat data [10]. More specifically, (multi)sensor data fusion problem aims at assessing the situation awareness of possible threats and understanding their relevance in the respective scenario. This can be done in many ways, such as: using sensor data deriving from radar, passive electronic supports, infrared identification-friend-foe sensors, electro-optic image sensors, etc. These kinds of data are not inherently relational, but physical streams of (semi)unstructed data.

In the statistical literature and in the marketing research, scientists often face the task of reconstructing/imputing missing data. This problem is named in variety of ways, including data fusion, but it is important to point out that it faces a very different issue, more often concerning statistical matching techniques [20]. Denominations include statistical data fusion [11], file concatenation method [17], micro data set merging [4], cross-tabulation [6, 11], or in the marketing field, multi-source imputation [8, 19]. The real task, without referring to this plethora of names, is linking a number of datasets with the goal of accessing the variables that are not present in a single source. In contrast with the data fusion problem, these datasets are not the output of any matching process [5, 16]: they are not structurally reconciled hence do not share any identifier, if any subsample of tuples contains the same units.

## 5 Conclusions

This paper presented BP-fuse as a novel algorithm to solve conflicts in database fusion. The major result is the possibility to exploit the dependencies among the features to solve the conflicts. These dependencies are modeled in Bayesian networks that represent domain knowledge. Dependencies among the attributes and non-conflicting values are used in conjunction in a global perspective, to establish which values are more plausible in the result. Once the knowledge has been captured by the Bayesian network, it can be used independently of the data, in this sense BP-fuse is context independent but domain aware.

## References

1. Bernstein PA (2003) Applying model management to classical meta data problems. CIDR 2003:209–220
2. Bilke A, Bleiholder J, Böhm C, Draba K, Naumann F, Weis M (2005) Automatic data fusion with hummer. Proc VLDB
3. Bleiholder J, Naumann F (2008) Data fusion. ACM Comput Surv (CSUR)
4. Budd EC (1971) The creation of a microdata file for estimating the size distribution of income. Rev Income Wealth 17(4):317–333
5. Christen P (2012) Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer
6. Fujii T, van der Weide R (2011) Two-sample cross-tabulation
7. Galindo-Legaria C (1994) Outerjoins as disjunctions. In: SIGMOD conference
8. Gilula Z, McCulloch RE, Rossi PE (2006) A direct approach to data fusion. J Mark Res 43(1):73–83
9. Halevy AY (2001) Answering queries using views: a survey. VLDB J (4)
10. Hall DL (2004) Mathematical techniques in multisensor data fusion
11. Kamakura WA, Wedel M (1997) Statistical data fusion for cross-tabulation. J Mark Res 485–498
12. Koller D, Friedman N (2009) Probabilistic graphical models. The MIT Press
13. Pearl J, Russel S (2011) Bayesian networks
14. Raghavan S, Garcia-Molina H (2001) Integrating diverse information management systems: a brief survey. IEEE Data Eng Bull 24(4):44–52
15. Rahm E, Do HH (2000) Data cleaning: problems and current approaches. IEEE Data Eng Bull 23(4):3–13
16. Rässler S (2012) Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches, vol 168. Springer
17. Rubin DB (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. J Bus Econ Stat 4(1):87–94
18. Ullman JD (1997) Information integration using logical views. In Database theory ICDT'97. Springer, pp 19–40
19. Van der Puttan P, Kok JN, Gupta A (2002) Data fusion through statistical matching. Alfred P. Sloan School of Management, Massachusetts Institute of Technology
20. Vantaggi B (2008) Statistical matching of multiple sources: a look through coherence. Int J Approximate Reasoning 49(3):701–711
21. Yan L, Tamer M (1999) Conflict tolerant queries in aurora. In: CoopIS. IEEE Computer Society, pp 279–290

# The Qualitative Characteristics of Combining Evidence with Discounting

**Alexander Lepskiy**

**Abstract** The qualitative characteristics of the combining evidence with the help of Dempster's rule with discounting is studied in this paper in the framework of Dempster-Shafer theory. The discount coefficient (discounting rate) characterizes the reliability of information source. The conflict between evidence and change of ignorance after applying combining rule are considered in this paper as important characteristics of quality of combining. The quantity of ignorance is estimated with the help of linear imprecision index. The set of crisp and fuzzy discounting rates for which the value of ignorance after combining does not increases is described.

**Keywords** Belief functions · Discount method · Imprecise index

## 1 Introduction

The study of combining rules of evidence occupies an important place in the belief functions theory. A combining rule puts in correspondence to two or more evidences the one evidence. Dempster's rule [4] was the first from combining rules. The review of some popular combining rules can be found in [10]. There is no combining rule which give a plausible aggregation of information in all cases regardless of context. The prognostic quality of combining evidence is evaluated with the help of some characteristics. The reliability of sources of information, the conflict measure of evidence [7], the degree of independence of evidence are a priori characteristics of quality of combining. The amount of change of ignorance after the use of a combining rule is the most important a posteriori characteristic [8]. The amount of ignorance contained in evidence may be estimated with the help of imprecision indices [2]. The generalized Hartley's measure is an example of such index [5]. It is known, for example, that the amount of ignorance does not increase when used Dempster's rule for non-conflicting evidences. Dempster's rule can be considered as an optimistic rule in this sense [8]. On the contrary, Dubois and Prade's disjunctive consensus

A. Lepskiy (✉)
Higher School of Economics, 20 Myasnitskaya Ulitsa, 101000 Moscow, Russia
e-mail: alex.lepskiy@gmail.com

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_39

rule [6] has a pessimistic character in the sense that amount of ignorance does not decrease after applying such a rule.

The discount method is one of the approaches where the reliability of information source is taken into account. This method was proposed by Shafer [11]. The discount coefficient (discounting rate) characterizes the reliability of information source. The discount method with Dempster's rule may be pessimistic rule or optimistic rule in depending on the values of discounting rates. The generalizations of the discount method were considered in several papers. In particular, Smets [12] introduced a family of combination rules known as $\alpha$-junctions. Pichon and Denoeux [9] have established the link between the parameter of $\alpha$-junction and reliability of information sources.

In this paper we will find conditions on the discount rates for which the amount of ignorance after applying Dempster's rule is not increased, i.e. this rule will be still optimistic in spite of unreliable information sources. This problem is solved in general case of conflicting evidences and crisp discounting rates as well as in the case of non-conflicting evidences and fuzzy discounting rates. In addition, the problem of finding such discount rates for which a conflict of evidence will not be greater than a certain threshold and the quality of ignorance after the combination will not increase is formulated and solved.

## 2 Belief Function Basics

Let $X$ be a finite universal set and $2^X$ is a set of all subsets of $X$. We consider the belief function [11] $g : 2^X \to [0, 1]$. The value $g(A)$, $A \in 2^X$, is interpreted as a degree of confidence that the true alternative of $X$ belongs to set $A$. A belief function $g$ is defined with the help of so called mass function $m_g : 2^X \to [0, 1]$ that satisfy the conditions [11]: $m_g(\emptyset) = 0$, $\sum_{A\subseteq X} m_g(A) = 1$. Then $g(A) = \sum_{B:\, B\subseteq A} m_g(B)$. Let the set of all belief functions on $2^X$ be denoted by $Bel(X)$. The belief function $g \in Bel(X)$ may be represented with the help of so called categorical belief functions $\eta_{\langle B\rangle}(A) = \begin{cases} 1, & B \subseteq A, \\ 0, & B \not\subseteq A, \end{cases}$ $A \subseteq X$, $B \neq \emptyset$. Then $g = \sum_{B\in 2^X\backslash\{\emptyset\}} m_g(B)\eta_{\langle B\rangle}$. The subset $A \in 2^X$ is called a focal element if $m(A) > 0$. Let $\mathcal{A}$ be a set of all focal elements. The pair $F = (\mathcal{A}, m)$ is called a body of evidence. We will denote through $\mathcal{A}(g)$ and $F(g)$ the set of all focal elements and the body of evidence correspondingly related with the belief function $g$. Let us have two bodies of evidence $F(g_1) = (\mathcal{A}(g_1), m_{g_1})$ and $F(g_2) = (\mathcal{A}(g_2), m_{g_2})$ which related with the belief functions $g_1, g_2 \in Bel(X)$. For example, it can be evidences which were received from two information sources. Then the task of combining of these two evidence in one evidence with the help of some operator $\varphi : Bel^2(X) \to Bel(X)$, $g = \varphi(g_1, g_2)$, is an actual problem. Dempster's rule was the first from combining rules. This rule was introduced in [4] and generalized in [11] for combining arbitrary independent evidence. This rule is defined as $g = \varphi_D(g_1, g_2) = \sum_{A\in 2^X\backslash\{\emptyset\}} m_g(A)\eta_{\langle A\rangle}$, where

$$m_g(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_{g_1}(B) m_{g_2}(C), \quad A \neq \emptyset, \quad m_g(\emptyset) = 0, \tag{1}$$

$$K = K(g_1, g_2) = \sum_{B \cap C = \emptyset} m_{g_1}(B) m_{g_2}(C).$$

The value $K(g_1, g_2)$ characterizes the amount of conflict in two information sources which determined with the help of bodies of evidence $F(g_1)$ and $F(g_2)$. If $K(g_1, g_2) = 1$ then it means that information sources are absolutely conflict and Dempster's rule cannot be applied. The discounting of mass function was introduced by Shafer [11] for accounting of reliability of information. The main idea consists in the use of coefficient $\alpha \in [0, 1]$ for discounting of mass function:

$$m^{\alpha}(A) = (1-\alpha)m(A), \quad A \neq X, \quad m^{\alpha}(X) = \alpha + (1-\alpha)m(X). \tag{2}$$

The coefficient $\alpha$ is called the discounting rate. The discounting rate characterized the degree of reliability of information. If $\alpha = 0$ then it means that information source is absolutely reliable. If $\alpha = 1$ then it means that information source is absolutely non-reliable. Dempster's rule (1) applies after discounting of mass functions of two evidences in general with different discounting rates.

The following Dubois and Prade's disjunctive consensus rule is a dual to Dempster's rule [6]: $g = \varphi_{DP}(g_1, g_2) = \sum_{A \in 2^X \setminus \{\emptyset\}} m_g(A) \eta_{\langle A \rangle}$, where $m_g(A) = \sum_{B \cup C = A} m_{g_1}(B) m_{g_2}(C)$, $A \in 2^X$.

## 3 Estimation of Ignorance Associated with the Belief Function

Let us have source of information and this information is described by a belief function $g \in Bel(X)$. The belief function $g$ defines the information with some degree of uncertainty. There are few approaches to definition of uncertainty measure in the evidence theory. We will follow the approach described in work [2]. This approach based on the notion of imprecision index.

Let us know only that true alternative belong to the non empty set $B \subseteq X$. This situation may be described with the help of categorical belief function $\eta_{\langle B \rangle}(A)$, $A \subseteq X$, which gives the lower probability of an event $x \in A$. The degree of uncertainty of such function is described by the well-known Hartley measure $H(\eta_{\langle B \rangle}) = \log_2 |B|$, which characterized the degree of information uncertainty about belonging of true alternative to set $B \subseteq X$.

The following construction is a generalization of above situation. Let $g = \sum_{B \in 2^X} m_g(B) \eta_{\langle B \rangle} \in Bel(X)$. Then the generalized Hartley measure [5] from $g$ is defined as $GH(g) = \sum_{B \in 2^X \setminus \{\emptyset\}} m_g(B) \log_2 |B|$. The generalized Hartley measure is an example of the following general notion.

**Definition 1** [2]. A functional $f : Bel(X) \to [0,1]$ is called imprecision index if the following conditions are fulfilled: (1) if $g$ be a probability measure then $f(g) = 0$; (2) $f(g_1) \geq f(g_2)$ for all $g_1, g_2 \in Bel(X)$ such that $g_1 \leq g_2$ (i.e. $g_1(A) \leq g_2(A)$ for all $A \in 2^X$); (3) $f\left(\eta_{\langle X\rangle}\right) = 1$.

An imprecision index $f$ on $Bel(X)$ is called linear if for any linear combination $\sum_{j=1}^{k} \alpha_j g_j \in Bel(X)$, $\alpha_j \in \mathbb{R}$, $g_j \in Bel(X)$, $j = 1, \ldots, k$, we have $f\left(\sum_{j=1}^{k} \alpha_j g_j\right) = \sum_{j=1}^{k} \alpha_j f\left(g_j\right)$.

The different representations of imprecision index were found in [2]. In this paper we will use the following representation.

**Proposition 1** *A functional $f : Bel(X) \to [0,1]$ is a linear imprecision index on $Bel(X)$ iff $f(g) = \sum_{B\in 2^X\setminus\{\emptyset\}} m_g(B)\mu_f(B)$, where set function $\mu_f$ satisfies the conditions: (1) $\mu_f(\{x\}) = 0$ for all $x \in X$; (2) $\mu_f(X) = 1$; (3) $\sum_{B:A\subseteq B} (-1)^{|B\setminus A|}\mu_f(B) \leq 0$ for all $A \neq \emptyset, X$.*

## 4 Change of Ignorance After Combining with the Crisp Discount Rates

Assume that we have two information sources which are defined by the bodies of evidence $F(g_1) = (\mathcal{A}(g_1), m_{g_1})$ and $F(g_2) = (\mathcal{A}(g_2), m_{g_2})$ correspondingly and which related with the belief functions $g_1, g_2 \in Bel(X)$. If we apply some combining rule $\varphi$ to the pair of belief functions $g_1, g_2 \in Bel(X)$ then we get a new belief function $g = \varphi(g_1, g_2)$. We have a question about changing of the amount of ignorance after applying combining rule $\varphi$. We will estimate the quantity of ignorance with the help of imprecision index $f$.

**Definition 2** A combining rule $\varphi$ is called optimistic (pessimistic) rule with respect to imprecision index $f$, if $f(g) \leq \min\limits_{i\in 1,2} f(g_i)$ $(f(g) \geq \max\limits_{i\in 1,2} f(g_i))$ for all $g_1, g_2 \in Bel(X)$.

In other words, the optimistic rule does not increase the amount of ignorance, but the pessimistic rule does not decrease the amount of ignorance. It is known [6, 8] that Dempster's rule is an optimistic rule with respect to any linear imprecision index, but Dubois and Prade's disjunctive consensus rule is a pessimistic rule.

Now we investigate on pessimism-optimism Dempster's rule with discounting. Let $g_1 = \sum_{A\in 2^X\setminus\{\emptyset\}} m_{g_1}(A)\eta_{\langle A\rangle}$ and $g_2 = \sum_{A\in 2^X\setminus\{\emptyset\}} m_{g_2}(A)\eta_{\langle A\rangle}$. Each of two information sources has its own reliability (discount rate) $\alpha, \beta \in [0,1]$ correspondingly in the sense of discounting method (2). We obtain two new belief functions taking into account discount rates:

$$g_1^{(\alpha)} = \sum\nolimits_{A\in 2^X\setminus\{\emptyset\}} m_{g_1}^{(\alpha)}(A)\eta_{\langle A\rangle}, \quad g_2^{(\beta)} = \sum\nolimits_{B\in 2^X\setminus\{\emptyset\}} m_{g_2}^{(\beta)}(B)\eta_{\langle B\rangle},$$

where $m_{g_1}^{(\alpha)}(A) = (1-\alpha)m_{g_1}(A), A \neq X, m_{g_1}^{(\alpha)}(X) = \alpha + (1-\alpha)m_{g_1}(X)$ and $m_{g_2}^{(\beta)}$ calculated similarly. We note that

$$g_1^{(\alpha)} = \sum\nolimits_{A \in 2^X \setminus \{\emptyset\}} m_{g_1}(A)\eta_{\langle A \rangle}^{(\alpha)}, \quad g_2^{(\beta)} = \sum\nolimits_{B \in 2^X \setminus \{\emptyset\}} m_{g_2}(B)\eta_{\langle B \rangle}^{(\beta)}, \tag{3}$$

where $\eta_{\langle A \rangle}^{(\alpha)} = (1-\alpha)\eta_{\langle A \rangle} + \alpha\eta_{\langle X \rangle}$ and $\eta_{\langle B \rangle}^{(\beta)}$ calculated similarly. We assume that evidences $F(g_1)$ and $F(g_2)$ are non-conflicting, i.e. $K = K(g_1, g_2) = 0$. Then $K\left(g_1^{(\alpha)}, g_2^{(\beta)}\right) = 0$. If we apply Dempster's rule $\varphi_D$ to the pair $g_1^{(\alpha)}, g_2^{(\beta)}$ of belief functions then we get a new belief function $g_{\alpha,\beta} = \varphi_D(g_1^{(\alpha)}, g_2^{(\beta)})$. Dempster's rule $\varphi_D(g_1^{(\alpha)}, g_2^{(\beta)})$ is a linear rule for every argument for non-conflicting evidences. Therefore we get from representations (3)

$$\varphi_D(g_1^{(\alpha)}, g_2^{(\beta)}) = \sum_{A \in \mathcal{A}(g_1)} \sum_{B \in \mathcal{A}(g_2)} m_{g_1}(A)m_{g_2}(B)\varphi_D\left(\eta_{\langle A \rangle}^{(\alpha)}, \eta_{\langle B \rangle}^{(\beta)}\right).$$

We have $A \cap B \neq \emptyset$ for every pair $A \in \mathcal{A}(g_1)$, $B \in \mathcal{A}(g_2)$ in case of non-conflicting evidences. Consequently we get

$$\varphi_D\left(\eta_{\langle A \rangle}^{(\alpha)}, \eta_{\langle B \rangle}^{(\beta)}\right) = (1-\alpha)(1-\beta)\eta_{\langle A \cap B \rangle} + (1-\alpha)\beta\eta_{\langle A \rangle} + \alpha(1-\beta)\eta_{\langle B \rangle} + \alpha\beta\eta_{\langle X \rangle}.$$

Therefore, a new belief function $g_{\alpha,\beta}$ has the following expression through initial functions $g_1, g_2 \in Bel(X)$ and the belief function $g = \varphi_D(g_1, g_2)$ obtained without discounting

$$g_{\alpha,\beta} = \varphi_D(g_1^{(\alpha)}, g_2^{(\beta)}) = (1-\alpha)(1-\beta)g + (1-\alpha)\beta g_1 + (1-\beta)\alpha g_2 + \alpha\beta\eta_{\langle X \rangle}. \tag{4}$$

We have a question about changing of the amount of ignorance after applying Dempster's rule with discounting. We will estimate the quantity of ignorance with the help of linear imprecision index $f$. Dempster's rule is an optimistic rule (i.e. $f(g) \leq \min\limits_i f(g_i)$,) for non-conflicting and reliable information sources ($\alpha, \beta = 0$) with respect to any linear imprecision index. If we use non-reliable information sources ($\alpha, \beta \neq 0$) then imprecision index $f(g_{\alpha,\beta})$ of new belief function $g_{\alpha,\beta}$ could be greater than imprecision indices of initial functions $f(g_i)$, $i = 1, 2$. We will find the conditions on discounting rates for which the amount of ignorance will not increase after applying Dempster's rule with discounting. We obtain from (4) with account of linearity of index $f$ and normalization condition $f(\eta_{\langle X \rangle}) = 1$ that

$$f(g_{\alpha,\beta}) = (1-\alpha)(1-\beta)f(g) + (1-\alpha)\beta f(g_1) + (1-\beta)\alpha f(g_2) + \alpha\beta.$$

The function $f(g_{\alpha,\beta})$ can be rewritten in the form

$$f(g_{\alpha,\beta}) = f(g) + \alpha\Delta_2 + \beta\Delta_1 + \alpha\beta(\Delta - \Delta_1 - \Delta_2), \tag{5}$$

where $\Delta_i = f(g_i) - f(g)$, $i = 1, 2$ is a changing of ignorance of $i$-th information source after applying Dempster's rule (without of discounting), $\Delta = 1 - f(g)$. Note that we have $\Delta_i \geq 0$, $i = 1, 2$ in any non-conflicting case and we have $\Delta \geq \Delta_i$, $i = 1, 2$ in any case. Then the condition $f(g_{\alpha,\beta}) \leq f(g_i)$, $i = 1, 2$ is equivalent to inequality

$$\alpha\Delta_2 + \beta\Delta_1 + \alpha\beta(\Delta - \Delta_1 - \Delta_2) \leq \min\{\Delta_1, \Delta_2\}. \tag{6}$$

Let $Ign_0 = Ign_0(g_1, g_2)$ be a set of all pair $(\alpha, \beta) \in [0, 1]^2$ which satisfy inequality (6) for given belief functions $g_1, g_2 \in Bel(X)$.

We have the following result in the general case of conflicting evidence (i.e. $K = K(g_1, g_2) \neq 0$).

**Proposition 2** *Dempster's rule with discounting* $(\alpha, \beta) \in [0, 1]^2$ *is optimistic rule with respect to linear imprecision index* $f$ *(i.e.* $f(g_{\alpha,\beta}) \leq \min_i f(g_i)$*) iff*

$$\alpha\Delta_2 + \beta\Delta_1 + \alpha\beta(\Delta - \Delta_1 - \Delta_2) \leq (1 - (1 - \alpha)(1 - \beta)K) \min\{\Delta_1, \Delta_2\}. \tag{7}$$

Let $Ign_K = Ign_K(g_1, g_2)$ be a set of all pair $(\alpha, \beta) \in [0, 1]^2$, which satisfy inequality (7) for given belief functions $g_1, g_2 \in Bel(X)$, which have conflict $K = K(g_1, g_2)$. It is easy to see from (7) that $Ign_{K'} \subseteq Ign_{K''} \subseteq Ign_0$, if $K' \geq K''$ under condition $\Delta_i = f(g_i) - f(g) \geq 0, i = 1, 2$.

The value of conflict after discounting is equal $K_{\alpha,\beta} = K\left(g_1^{(\alpha)}, g_2^{(\beta)}\right) = (1 - \alpha)(1 - \beta)K$. If the discount rates are increased then the value of conflict between the evidence is decreased. The problem of description of all pair $(\alpha, \beta) \in [0, 1]^2$ for given belief functions $g_1, g_2 \in Bel(X)$ for which the conflict $K_{\alpha,\beta}$ is not greater some threshold value $K_{\max} \leq K$ (i.e. $K_{\alpha,\beta} = (1 - \alpha)(1 - \beta)K \leq K_{\max}$) can be formulated. We denote this set through $Confl_K(K_{\max})$.

The problem of description of reliability of information sources (discounting rates) for which the aggregation with the help of Dempster's rule will not lead to an increase of ignorance ($(\alpha, \beta) \in Ign_K$) but a conflict will not be great ($(\alpha, \beta) \in Confl_K(K_{\max})$) is an actual problem. This set is defined as $Ign_K \cap Confl_K(K_{\max})$.

Now the problem of finding of points-reliabilities $(\alpha, \beta) \in Ign_K \cap Confl_K(K_{\max})$ for which the imprecision index $f(g_{\alpha,\beta})$ after combining will be minimal can be formulated:

$$f(g_{\alpha,\beta}) \to \min, \quad (\alpha, \beta) \in Ign_K \cap Confl_K(K_{\max}). \tag{8}$$

This problem is an actual if we have several pairs of conflicting information sources with different reliabilities. We must choose the best pair for combining. Note that the formulation of the problem (8) can be considered as an optimization problem of finding of combining rule from parametric family of rules $\left\{g_{\alpha,\beta}\right\}_{\alpha,\beta\in[0,1]}$, for which the ignorance will be minimal under the condition that the conflict is not greater some threshold value $K_{\max}$. The generalized statement of the problem is considered in [3].

## 5 Change of Ignorance After Combining with Fuzzy Discount Rates

Assume that reliabilities of information sources $\alpha$ and $\beta$ are not known precisely but we have a fuzzy numbers $\tilde{\alpha}$ and $\tilde{\beta}$. Then the imprecision index $f(g_{\tilde{\alpha},\tilde{\beta}})$ will be by a fuzzy number also and, for example, in case of non-conflicting evidence (see (5)) $f(g_{\tilde{\alpha},\tilde{\beta}})$ is equal

$$f(g_{\tilde{\alpha},\tilde{\beta}}) = f(g) + \tilde{\alpha}\Delta_2 + \tilde{\beta}\Delta_1 + \tilde{\alpha}\tilde{\beta}(\Delta - \Delta_1 - \Delta_2).$$

Then we can formulate the problem of finding of the fuzzy numbers $\tilde{\alpha}$ and $\tilde{\beta}$ for which $f(g_{\tilde{\alpha},\tilde{\beta}}) \leq_I f(g_i)$, $i = 1, 2$, where $\leq_I$ is a some relation of comparison of fuzzy numbers.

*Example* Let $\tilde{\alpha}$ and $\tilde{\beta}$ are by triangular fuzzy numbers of the form $\tilde{\alpha} = (\alpha - \delta, \alpha, \alpha + \delta)$ and $\tilde{\beta} = (\beta - \omega, \beta, \beta + \omega)$ correspondingly. We will use the method Adamo [1] for comparison of the fuzzy numbers $\tilde{u}$ and $\tilde{v}$. Let $\tilde{u}_\gamma = \{t | \mu_{\tilde{u}}(t) \geq \gamma\}$ be a $\gamma$-cut of fuzzy number $\tilde{u}$ with relationship function $\mu_{\tilde{u}}$ and $\tilde{u}_\gamma = [l_{\tilde{u}}(\gamma), r_{\tilde{u}}(\gamma)]$. The fuzzy number $\tilde{u}$ does not exceed the fuzzy number $\tilde{v}$ with respect to the method Adamo ($\tilde{u} \leq_A \tilde{v}$), if $r_{\tilde{u}}(\gamma) \leq r_{\tilde{v}}(\gamma)$ for given (fixed) level $\gamma \in (0, 1]$. The level $\gamma$ characterizes a measure of risk of the wrong decision. Then

$$f(g_{\tilde{\alpha},\tilde{\beta}}) \leq_I f(g_i) \Leftrightarrow r_{f(g_{\tilde{\alpha},\tilde{\beta}})}(\gamma) \leq \min_i f(g_i),$$

where $r_{f(g_{\tilde{\alpha},\tilde{\beta}})}(\gamma) = f(g) + r_{\tilde{\alpha}}(\gamma) v \Delta_2 + r_{\tilde{\beta}}(\gamma)\Delta_1 + r_{\tilde{\alpha}}(\gamma) r_{\tilde{\beta}}(\gamma)(\Delta - \Delta_1 - \Delta_2)$, $r_{\tilde{\alpha}}(\gamma) = \alpha + \delta(1 - \gamma)$, $r_{\tilde{\beta}}(\gamma) = \beta + \omega(1 - \gamma)$, $\gamma \in (0, 1]$.

## 6 Conclusion

The qualitative characteristics of the combining evidence with the help of Dempster's rule with discounting were studied in this paper in the framework of Dempster-Shafer theory. In particular we found conditions on the discount rates for which the amount of ignorance after applying Dempster's rule is not increased, i.e. this rule will be still optimistic in spite of unreliable information sources. This problem was solved in general case of conflicting evidences and crisp discounting rates as well as in the case of non-conflicting evidences and fuzzy discounting rates. In addition, the problem of finding such discount rates for which a conflict of evidence will not be greater than a certain threshold and the quality of ignorance after the combination will not increase was formulated and solved.

**Acknowledgments** The financial support from the Government of the Russian Federation within the framework of the implementation o f the 5–100 Programme Roadmap of the National Research University Higher School of Economics is acknowledged. This work was supported by the grant 14-07-00189 of RFBR (Russian Foundation for Basic Research).

## References

1. Adamo JM (1980) Fuzzy decision trees. Fuzzy Sets Syst 4:207–219
2. Bronevich A, Lepskiy A (2015) Imprecision indices: axiomatic, properties and applications. Int J Gen Syst 44(7–8):812–832
3. Bronevich A, Rozenberg I (2015) The choice of generalized dempstershafer rules for aggregating belief functions. Int J Approximate Reasoning 56-A:122–136
4. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38(2):325–339
5. Dubois D, Prade H (1985) A note on measures of specificity for fuzzy sets. Int J Gen Syst 10:279–283
6. Dubois D, Prade H (1986) A set-theoretic view of belief functions logical operations and approximation by fuzzy sets. Int J Gen Syst 12(3):193–226
7. Lepskiy A (2013) About relation between the measure of conflict and decreasing of ignorance in theory of evidence. In: Proceedings of the 8th conference of the European society for fuzzy logic and technology. Atlantis Press, Amsterdam, pp 355–362
8. Lepskiy A (2014) General schemes of combining rules and the quality characteristics of combining. In: BELIEF 2014, LNAI 8764. Springer, pp 29–38
9. Pichon F, Denoeux T (2009) Interpretation and computation of alpha-junctions for combining belief functions. In: 6th international symposium on imprecise probability: theories and applications (ISIPTA '09), Durham, UK
10. Sentz K, Ferson S (2002) Combination of evidence in dempster-shafer theory. In: Report SAND 2002-0835, Sandia Nat Labor
11. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
12. Smets Ph (1995) The canonical decomposition of weighted belief. In: International joint conference on artificial intelligence, pp 1896–1901

# Measuring the Dissimilarity Between the Distributions of Two Random Fuzzy Numbers

**María Asunción Lubiano, María Ángeles Gil, Beatriz Sinova, María Rosa Casals and María Teresa López**

**Abstract** In a previous paper the fuzzy characterizing function of a random fuzzy number was introduced as an extension of the moment generating function of a real-valued random variable. Properties of the fuzzy characterizing function have been examined, among them, the crucial one proving that it unequivocally determines the distribution of a random fuzzy number in a neighborhood of 0. This property suggests to consider the empirical fuzzy characterizing function as a tool to measure the dissimilarity between the distributions of two random fuzzy numbers, and its expected descriptive potentiality is illustrated by means of a real-life example.

## 1 Introduction

The formalization of random fuzzy numbers as Borel-measurable fuzzy number-valued mappings associated with a probability space, this one modeling a random experiment, allows us to properly refer to its induced distribution as well as to the independence of random fuzzy numbers. Nevertheless, although the existence of such an induced distribution is clear (and it can be easily determined in the sample case), there is not a sound general concept which enables us to develop some probabilistic and statistical results we have in the real-valued case, like the distribution function of a real-valued random variable. Moreover, there are not exact or approximated models widely applicable and realistic enough for the induced distribution.

M.A. Lubiano · M.Á. Gil (✉) · B. Sinova · M.R. Casals · M.T. López
Departamento de Estadística e I.O. y D.M., University of Oviedo, 33071 Oviedo, Spain
e-mail: magil@uniovi.es

M.A. Lubiano
e-mail: lubiano@uniovi.es

B. Sinova
e-mail: sinovabeatriz@uniovi.es

M.R. Casals
e-mail: rmcasals@uniovi.es

M.T. López
e-mail: mtlopez@uniovi.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_40

In Sinova et al. [9] a function characterizing the induced distribution of a random fuzzy number has been defined. This function aims to extend the moment generating function of a real-valued random variable (and, therefore, there are just a few distributions for which it does not exist) and it is based on the Aumann-type mean of a random fuzzy number. Since the extension preserves the convenient characterizing ability of the moment generating function, one can think of using it to measure to some extent whether the (induced) distributions of two random fuzzy numbers coincide or not. More concretely, we can consider to state a measure of the dissimilarity of such distributions.

This paper aims to empirically analyze the descriptive behaviour of this measure by means of a real-life example. The derived descriptive conclusions will be compared with some inferential ones which have been recently drawn. Some open problems will be finally proposed.

## 2 Preliminaries

Fuzzy sets, and particularly fuzzy numbers, are very suitable to cope with the imprecision of different real-life data, especially those coming from human thought and experience in variables like quality perception, satisfaction, opinion, etc.

**Definition 1** A mapping $\widetilde{U} : \mathbb{R} \to [0, 1]$ is said to be a (bounded) *fuzzy number* if its $\alpha$-levels

$$\widetilde{U}_\alpha = \begin{cases} \{x \in \mathbb{R} \,:\, \widetilde{U}(x) \geq \alpha\} & \text{if } \alpha \in (0, 1] \\ \text{cl}\{x \in \mathbb{R} \,:\, \widetilde{U}(x) > 0\} & \text{if } \alpha = 0 \end{cases}$$

(with cl denoting the topological closure) are nonempty compact intervals for all $\alpha \in [0, 1]$. The class of (bounded) fuzzy numbers will be denoted by $\mathcal{F}_c^*(\mathbb{R})$.

To deal with fuzzy numbers in this paper we should consider the extension of the sum and product by a scalar as well as that for the exponential function, which will be supposed to be based on Zadeh's extension principle [10] and coincides level-wise with the usual interval arithmetic and function image (see Nguyen [7]).

**Definition 2** Let $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c^*(\mathbb{R})$ and $\gamma \in \mathbb{R}$. The *sum* of $\widetilde{U}$ and $\widetilde{V}$ is the fuzzy number $\widetilde{U} + \widetilde{V}$ such that

$$(\widetilde{U} + \widetilde{V})_\alpha = \text{Minkowski sum of } \widetilde{U}_\alpha \text{ and } \widetilde{V}_\alpha = [\inf \widetilde{U}_\alpha + \inf \widetilde{V}_\alpha, \sup \widetilde{U}_\alpha + \sup \widetilde{V}_\alpha].$$

The *product* of $\widetilde{U}$ *by the scalar* $\gamma$ is the fuzzy number $\gamma \cdot \widetilde{U}$ such that

$$(\gamma \cdot \widetilde{U})_\alpha = \gamma \cdot \widetilde{U}_\alpha = \begin{cases} \left[\gamma \inf \widetilde{U}_\alpha, \gamma \sup \widetilde{U}_\alpha\right] & \text{if } \gamma \in [0, \infty) \\ \left[\gamma \sup \widetilde{U}_\alpha, \gamma \inf \widetilde{U}_\alpha\right] & \text{otherwise.} \end{cases}$$

The (induced) *image of $\widetilde{U}$ through the exponential function* is the fuzzy number $e^{\gamma\cdot\widetilde{U}}$ such that

$$(e^{\gamma\cdot\widetilde{U}})_\alpha = \begin{cases} \left[e^{\gamma \inf \widetilde{U}_\alpha}, e^{\gamma \sup \widetilde{U}_\alpha}\right] & \text{if } \gamma \in [0,\infty) \\ \left[e^{\gamma \sup \widetilde{U}_\alpha}, e^{\gamma \inf \widetilde{U}_\alpha}\right] & \text{otherwise.} \end{cases}$$

If a random experiment leads to data which can be suitably modeled in terms of fuzzy numbers, one should also properly model the random mechanism generating such data to analyze them in a rigorously established setting. The concept of random fuzzy number (or one-dimensional fuzzy random variable, as coined and introduced by Puri and Ralescu [8]) is an appropriate model to formalize a random mechanism associating with each experimental outcome a fuzzy number. That is, random fuzzy numbers are mainly addressed to deal with the 'ontic' view of experimental fuzzy data (see Couso and Dubois [1]).

**Definition 3** Let $\mathcal{K}_c(\mathbb{R})$ be the space of nonempty compact intervals. Given a probability space $(\Omega, \mathcal{A}, P)$, a *random fuzzy number* associated with it is a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R})$ such that for each $\alpha \in [0,1]$ the set-valued mapping $\mathcal{X}_\alpha : \Omega \to \mathcal{K}_c(\mathbb{R})$ (with $\mathcal{X}_\alpha(\omega) = \big(\mathcal{X}(\omega)\big)_\alpha$) is a compact random interval.

Equivalently, a *random fuzzy number* is a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R})$ such that it is Borel-measurable w.r.t. the Borel $\sigma$-field generated on $\mathcal{F}_c^*(\mathbb{R})$ by the topology induced by several different metrics, like the 2-norm distance

$$\rho_2(\widetilde{U},\widetilde{V}) = \sqrt{\frac{1}{2}\int_{[0,1]} \left(\left[\inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha\right]^2 + \left[\sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha\right]^2\right) d\alpha}$$

by Diamond and Kloeden [2].

As we have already pointed out, the assumed Borel-measurability of random fuzzy numbers in the second equivalent definition allows us to trivially induce the distribution (from $P$) of a random fuzzy number.

A relevant measure in summarizing such an induced distribution is the mean value, which has been defined by Puri and Ralescu [8]) as follows:

**Definition 4** Given a probability space $(\Omega, \mathcal{A}, P)$ and a random fuzzy number $\mathcal{X}$ associated with it, the *(population) Aumann-type mean value of $\mathcal{X}$* is the fuzzy number $\widetilde{E}(\mathcal{X})$, if it exists, such that for each $\alpha \in [0,1]$

$$\big(\widetilde{E}(\mathcal{X})\big)_\alpha = \left[E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)\right].$$

In particular, if one deals with a finite sample of observations from a random fuzzy number $\mathcal{X}$, say $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$, the corresponding *(sample) Aumann-type mean value* is the fuzzy number

$$\overline{\widetilde{\boldsymbol{x}}} = \frac{1}{n}\cdot(\widetilde{x}_1 + \cdots + \widetilde{x}_n).$$

On the basis of the Aumann-type mean value of a random fuzzy number, one can formally extend the notion of moment generating function of a real-valued random variable as follows (see Sinova et al. [9]):

**Definition 5** Given a probability space $(\Omega, \mathcal{A}, P)$ and a random fuzzy number $\mathcal{X}$ associated with it, the *(population) fuzzy characterizing function* of $\mathcal{X}$ is the mapping $\widetilde{M}_{\mathcal{X}}$ defined on a neighborhood of 0 that associates with each $t$ in the neighborhood the fuzzy number $\widetilde{M}_{\mathcal{X}}(t) = \widetilde{E}\left(e^{t\mathcal{X}}\right)$, if it exists. That is, for each $\alpha \in [0, 1]$

$$\left(\widetilde{M}_{\mathcal{X}}(t)\right)_\alpha = \begin{cases} \left[E(e^{t \inf \mathcal{X}_\alpha}), E(e^{t \sup \mathcal{X}_\alpha})\right] & \text{if } t \geq 0 \\ \left[E(e^{t \sup \mathcal{X}_\alpha}), E(e^{t \inf \mathcal{X}_\alpha})\right] & \text{otherwise.} \end{cases}$$

In particular, if one deals with a finite sample of observations from a random fuzzy number $\mathcal{X}$, say $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$, the corresponding *empirical fuzzy characterizing function* is the mapping $\widehat{\widetilde{M}_{\widetilde{\boldsymbol{x}}}}$ associating with each $t$ in a neighborhood of 0 the fuzzy number

$$\widehat{\widetilde{M}_{\widetilde{\boldsymbol{x}}}}(t) = \frac{1}{n} \cdot \left(e^{t\widetilde{x}_1} + \cdots + e^{t\widetilde{x}_n}\right).$$

As shown in [9], the fuzzy characterizing function preserves most of the properties of the moment generating one in the real-valued case, but the one associated with the moment generation. However, it keeps the crucial property of characterization of the induced distribution of the associated random element, so that if $\mathcal{X}$ and $\mathcal{Y}$ are two random fuzzy numbers for which the fuzzy characterizing functions exist and coincide in a neighborhood of 0, then $\mathcal{X}$ and $\mathcal{Y}$ should be equally distributed.

In the next section, we are going to take advantage of this characterizing skill to state a descriptive measure for the dissimilarity between the sample distributions of two random fuzzy numbers.

## 3 A Sample Measure for the Dissimilarity Between the Distributions of Two Fuzzy Datasets

This section aims to state an index for the dissimilarity between the distributions of two fuzzy datasets. Due to the characterizing property, and being inspired by ideas in some statistics for the homogeneity of distributions in the real-valued case (see, for instance, Meintanis [5], Mora and Mora-López [6], who also suggest the correction in contrast to the measure in Lubiano et al. [3]), it seems plausible to consider in the current setting a statistic based on distances between the sample fuzzy characterizing functions in a narrow neighborhood of 0.

In this way, for an arbitrarily fixed $\varepsilon > 0$:

**Definition 6** The *$\varepsilon$-sample dissimilarity between the distributions of samples* $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ *and* $\widetilde{\boldsymbol{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_m)$ is given by the index

$$\varrho_{n,m,\varepsilon}(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) = \frac{1}{\varepsilon}\sqrt{\frac{nm}{n+m}} \max_{t\in[-\varepsilon,\varepsilon]} \rho_2\left(\widehat{\widetilde{M}}_{\widetilde{\boldsymbol{x}}}(t), \widehat{\widetilde{M}}_{\widetilde{\boldsymbol{y}}}(t)\right).$$

In this section we are going to apply the preceding measure on a dataset from a real-life situation.

*Example* The nine items displayed in Table 1 have been drawn from the TIMSS/ PIRLS 2011 Student questionnaire. This questionnaire is conducted in many countries and it is to be responded by fourth grade students (nine to ten years old) in connection with some aspects about reading, math and science.

These nine items have been originally designed to be answered in accordance with a 4-point Likert scale (DISAGREE A LOT, DISAGREE A LITTLE, AGREE A LITTLE, AGREE A LOT).

Recently, the questionnaire form involving these nine items, along with a few more ones about students' support resources at home, has been adapted to allow also a fuzzy rating scale-based one (see Fig. 1 for Question $M.2$). For the full paper-and-pencil and computerized versions of the questionnaire, see http://bellman.ciencias.uniovi.es/SMIRE/FuzzyRatingScaleQuestionnaire-SanIgnacio.html.

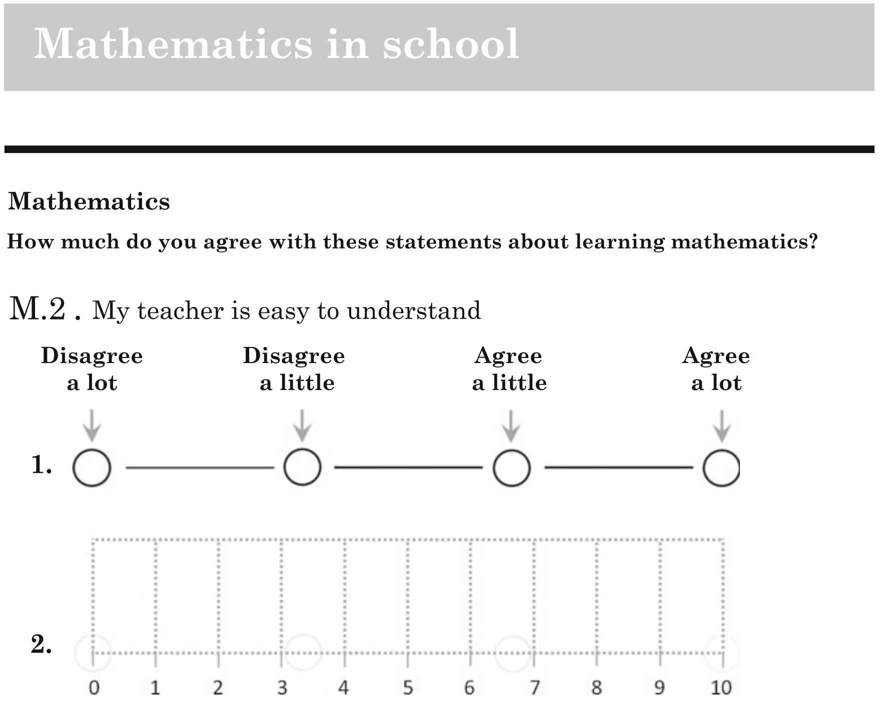
The fuzzy rating scale (see, e.g., [3, 4]) has been designed with reference interval [0, 10]. The adapted questionnaire has been conducted on 69 fourth grade students from Colegio San Ignacio (Oviedo-Asturias, Spain). The complete dataset can be found in the webpage containing the forms.

Now we are going to examine whether the fuzzy rating scale-based responses seem or not to be affected by respondents' sex, filled form version and the fact that respondents have or not an individual bedroom at home.

For this purpose, and for each of the three variables, we have first considered the (descriptive) dissimilarity index with $\varepsilon = 0.001, 0.01$ and $0.1$ (a deeper and exhaustive discussion about the choice of $\varepsilon$ should be developed in the future). Secondly, as an alternative (albeit inferential) way to discuss such an influence, we

**Table 1** Items selected from the TIMSS-PIRLS 2011 Student Questionnaire

| | *Reading in school* |
|---|---|
| $R.1$ | I like to read things that make me think |
| $R.2$ | I learn a lot from reading |
| $R.3$ | Reading is harder for me than any other subject |
| | *Mathematics in school* |
| $M.1$ | I like mathematics |
| $M.2$ | My teacher is easy to understand |
| $M.3$ | Mathematics is harder for me than any other subject |
| | *Science in school* |
| $S.1$ | My teacher taught me to discover science in daily life |
| $S.2$ | I read about in my spare time |
| $S.3$ | Science is harder for me than any other subject |



Mathematics in school

Mathematics

How much do you agree with these statements about learning mathematics?

M.2 . My teacher is easy to understand

Disagree a lot
Disagree a little
Agree a little
Agree a lot
1.
2.
0
1
2
3
4
5
6
7
8
9
10

**Fig. 1** Example of the double-response form to an item

have considered tests in Lubiano et al. [4] for the two-sample equality of independent means and compute the associated $p$-values when the chosen metric is $\rho_2$. Tables 2, 3 and 4 gather the outputs for the descriptive and inferential analyses.

**Table 2** $\varepsilon$-sample dissimilarity between the distributions of girls' and boys' samples for $\varepsilon = 0.001, 0.01, 0.1$ and $\rho_2$-based testing $p$-values for the equality of means

| Item | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.001$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.01$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.1$) | $\rho_2$ two-sample test $p$-values |
|---|---|---|---|---|
| $R.1$ | 0.3874 | 0.4056 | 0.665 | 0.502 |
| $R.2$ | 0.2397 | 0.2544 | 0.4759 | 0.702 |
| $R.3$ | 0.6087 | 0.6416 | 1.1206 | 0.425 |
| $M.1$ | 1.2692 | 1.3337 | 2.2487 | 0.049 |
| $M.2$ | 0.3713 | 0.39 | 0.658 | 0.574 |
| $M.3$ | 0.6207 | 0.6469 | 1.0211 | 0.49 |
| $S.1$ | 0.6784 | 0.7145 | 1.2232 | 0.275 |
| $S.2$ | 0.2754 | 0.2942 | 0.5738 | 0.687 |
| $S.3$ | 0.4223 | 0.4394 | 0.6851 | 0.606 |

**Table 3** $\varepsilon$-sample dissimilarity between the distributions of paper-and-pencil respondents' sample and computerized respondents' sample for $\varepsilon = 0.001, 0.01, 0.1$ and $\rho_2$-based testing $p$-values for the equality of means

| Item | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.001$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.01$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.1$) | $\rho_2$ two-sample test $p$-values |
|---|---|---|---|---|
| $R.1$ | 1.0148 | 1.0606 | 1.678 | 0.065 |
| $R.2$ | 1.1045 | 1.1724 | 2.1556 | 0.029 |
| $R.3$ | 0.7244 | 0.7497 | 1.0904 | 0.366 |
| $M.1$ | 0.8622 | 0.9008 | 1.4245 | 0.176 |
| $M.2$ | 1.3347 | 1.4103 | 2.5025 | 0.01 |
| $M.3$ | 1.5316 | 1.6148 | 2.8161 | 0.062 |
| $S.1$ | 1.5403 | 1.6122 | 2.6124 | 0.016 |
| $S.2$ | 0.6827 | 0.7058 | 0.9985 | 0.292 |
| $S.3$ | 1.5221 | 1.5978 | 2.664 | 0.042 |

**Table 4** $\varepsilon$-sample dissimilarity between the distributions of respondents' sample with individual bedroom and respondents' sample with shared bedroom for $\varepsilon = 0.001, 0.01, 0.1$ and $\rho_2$-based testing $p$-values for the equality of means

| Item | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.001$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.01$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = 0.1$) | $\rho_2$ two-sample test $p$-values |
|---|---|---|---|---|
| $R.1$ | 0.5859 | 0.6277 | 1.2509 | 0.294 |
| $R.2$ | 1.2238 | 1.3036 | 2.4909 | 0.013 |
| $R.3$ | 0.4755 | 0.4983 | 0.8005 | 0.543 |
| $M.1$ | 0.9392 | 0.9919 | 1.7486 | 0.188 |
| $M.2$ | 0.3153 | 0.3365 | 0.6604 | 0.685 |
| $M.3$ | 0.6548 | 0.6987 | 1.3606 | 0.46 |
| $S.1$ | 0.2659 | 0.2746 | 0.394 | 0.772 |
| $S.2$ | 0.5868 | 0.6063 | 0.8561 | 0.373 |
| $S.3$ | 0.8058 | 0.859 | 1.6633 | 0.366 |

As an attempt to analyze the coherence between the descriptive dissimilarity and the inferential testing for the equality of means outputs, we have computed Pearson's correlation coefficient $r$ between both series of outputs. In connection with sex we have that $r = -0.9567$ (if $\varepsilon = 0.001$), $r = -0.9574$ (if $\varepsilon = 0.01$), and $r = -0.9572$ (if $\varepsilon = 0.1$).

In connection with the filled format we have that $r = -0.8269$ (if $\varepsilon = 0.001$), $r = -0.8331$ (if $\varepsilon = 0.01$), and $r = -0.8664$ (if $\varepsilon = 0.1$). In connection with bedroom type for respondents we have that $r = -0.9437$ (if $\varepsilon = 0.001$), $r = -0.9426$ (if $\varepsilon = 0.01$), and $r = -0.9145$ (if $\varepsilon = 0.1$).

Consequently, there is a high linear relationship between both tools. Notice that the correlation coefficient is not expected to be exactly equal to $-1$, not only because we are using samples and linearity could be a restrictive assumption, but also because the dissimilarity index is related to the whole distribution whereas $p$-values concern only their means.

## 4 Conclusions and Future Directions

By looking at the outputs in Table 2, one can conclude both descriptively (through the dissimilarity measure) and inferentially (through the $p$-value) that sex affects the liking for mathematics (related to item $M.1$). Actually, $M.1$ is the only item among the 9 in the adapted questionnaire for which $\varrho_{n,m,0.001} > 1$ and the $p$-value is lower than 0.05.

By looking at the outputs in Table 3, one can conclude that the version form affects (to a rather great extent) the response to items $R.1$, $R.2$, $M.2$, $M.3$, $S.1$ and $S.3$, for which $\varrho_{n,m,0.001} > 1$ and the $p$-value is always lower or much lower than 0.07.

By looking at the outputs in Table 4, one can conclude that having or not an individual bedroom at home affects students' learning from reading (related to item $R.2$), for which $\varrho_{n,m,0.001} > 1$ and the $p$-value is lower than 0.02.

On the other hand, the measure in this paper has been simply applied for descriptive purposes. Consequently, we cannot attempt to interpret the significance of the dissimilarity measure. It would be desirable to consider this measure in the near future to develop inferential methods (more concretely, for testing hypothesis about the homogeneity of the population distributions of two random fuzzy numbers).

**Acknowledgments** Authors are grateful to Colegio San Ignacio in Oviedo-Asturias (Spain) for allowing us to collect the data in the real-life example. The research in this paper has been partially supported by/benefited from Principality of Asturias Grant GRUPIN14-101, and the Spanish Ministry of Economy and Competitiveness Grants MTM2015-63971-P and MTM2013-44212-P. Their financial support is gratefully acknowledged.

## References

1. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: Ontic vs. epistemic views. Int J Appr Reas 55(7):1502–1518
2. Diamond P, Kloeden P (1999) Metric spaces of fuzzy sets. Fuzzy Sets Syst 100:63–71
3. Lubiano MA, De la Rosa de Sáa S, Montenegro M, Sinova B, Gil, MA (2016) Descriptive analysis of responses to items in questionnaires. Why not using a fuzzy rating scale? Inform Sci 360:131–148
4. Lubiano MA, Montenegro M, Sinova B, De la Rosa de Sáa S, Gil MA (2016) Hypothesis testing for means in connection with fuzzy rating scale-based data: algorithms and applications. Eur J Oper Res 251:918–929
5. Meintanis SG (2007) A KolmogorovSmirnov type test for skew normal distributions based on the empirical moment generating function. J Stat Plan Infer 137:2681–2688
6. Mora J, Mora-López L (2010) Comparing distributions with bootstrap techniques: an application to global solar radiation. Math Comp Simul 81:811–819

7. Nguyen HT (1978) A note on the extension principle for fuzzy sets. J Math Anal Appl 64:369–380
8. Puri ML, Ralescu DA (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
9. Sinova B, Casals MR, Gil MA, Lubiano MA (2015) The fuzzy characterizing function of the distribution of a random fuzzy number. Appl Math Model 39(14):4044–4056
10. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning, Part 1. Inform Sci 8:199–249; Part 2. Inform Sci 8:301–353; Part 3. Inform Sci 8:43–80

# An Empirical Analysis of the Coherence Between Fuzzy Rating Scale- and Likert Scale-Based Responses to Questionnaires

María Asunción Lubiano, Antonia Salas, Sara De la Rosa de Sáa,
Manuel Montenegro and María Ángeles Gil

**Abstract** In dealing with questionnaires concerning satisfaction, quality perception, attitude, judgement, etc., the fuzzy rating scale has been introduced as a flexible way to respond to questionnaires' items. Designs for this type of questionnaires are often based on Likert scales. This paper aims to examine three different real-life examples in which respondents have been allowed to doubly answer: in accordance with either a fuzzy rating scale or a Likert one. By considering a minimum distance-based criterion, each of the fuzzy rating scale answers is associated with one of the Likert scale labels. The percentages of coincidences between the two responses in the double answer are computed by the criterion-based association. Some empirical conclusions are drawn from the computation of such percentages.

## 1 Introduction

In designing questionnaires concerning variables which cannot be measured by means of exact numerical values but can be graded to some extent (as it happens with satisfaction, quality perception, agreement level, and so on), commonly employed

M.A. Lubiano · A. Salas (✉) · S. De la Rosa de Sáa · M. Montenegro · M.Á. Gil
Departamento de Estadística e I.O. y D.M., University of Oviedo,
33071 Oviedo, Spain
e-mail: antonia@uniovi.es

M.A. Lubiano
e-mail: lubiano@uniovi.es

S. De la Rosa de Sáa
e-mail: rosasara@uniovi.es

M. Montenegro
e-mail: mmontenegro@uniovi.es

M.Á. Gil
e-mail: magil@uniovi.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_41

scales are Likert ones. Items in Likert scale-based questionnaires are responded by choosing among a list of a few pre-specified answers the one that best represents respondent's valuation, rating, opinion, etc. Likert scale-based answers can be usually ordered with respect to a certain criterion (say degree of satisfaction, degree of goodness, degree of agreement, etc.).

Hesketh et al. [5] (see also Hesketh and Hesketh [4]) proposed the so-called fuzzy rating scale to allow a complete freedom and expressiveness in responding, without respondents being constrained to choose among a few pre-specified responses. By drawing the fuzzy number that best represents respondent's valuation, the fuzzy rating scale captures the logical imprecision associated with such variables and allows us to have a rich continuous scale of measurement. In this way, the fuzzy rating scale somehow combines the power of the fuzzy linguistic scales with the power of visual analogue scales.

In previous papers, responses to items in synthetic and real-life questionnaires based both on Likert and fuzzy rating scales have been empirically compared by means of different statistical tools (see, for instance, De la Rosa de Sáa et al. [1], Gil et al. [3] and Lubiano et al. [7]).

Since responses in accordance with the two scales are collected in a linked way (i.e., respondents supply a double answer), one question that arises is whether or not respondents follow a kind of systematic classification of the fuzzy rating scale-based responses into classes that could be identified with Likert's possible answers.

This paper aims to examine this question by analyzing three real-life examples involving questionnaires with double response type items. For this purpose a criterion based on a distance between Likert and fuzzy responses (actually, between numerically encoded Likert and fuzzy responses) is applied. This analysis evidences that the coincidences between the expected Likert response and the one really chosen are high, but up to 90 %. This suggests that in assigning fuzzy rating scale-based responses people behave in a very free way, without trying to exactly follow a kind of fuzzy linguistic description of a Likert response. Furthermore, this fact corroborates to some extent that, as it has been frequently pointed out in the literature, the usual numerical encoding of Likert responses is not appropriate enough.

## 2 Preliminaries

Fuzzy numbers are often considered to express imprecise data because of their ability and power to precisiate the imprecision and to be mathematically handled.

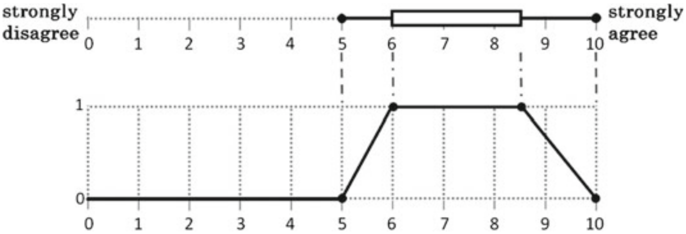
**Definition 1** A mapping $\widetilde{U} : \mathbb{R} \to [0, 1]$ is said to be a (bounded) *fuzzy number* if its $\alpha$-levels

$$\widetilde{U}_\alpha = \begin{cases} \{x \in \mathbb{R} : \widetilde{U}(x) \geq \alpha\} & \text{if } \alpha \in (0, 1] \\ \text{cl}\{x \in \mathbb{R} : \widetilde{U}(x) > 0\} & \text{if } \alpha = 0 \end{cases}$$

(with cl denoting the topological closure) are nonempty compact intervals for all $\alpha \in [0, 1]$. The class of (bounded) fuzzy numbers will be denoted by $\mathcal{F}_c^*(\mathbb{R})$.

In accordance with Hesketh et al. [5] (see also Hesketh and Hesketh [4]), the guideline for the use of fuzzy numbers through the so-called *fuzzy rating scale* is the following:

1. A reference bounded interval/segment $[a, b]$ is first considered. This is often chosen to be $[0, 10]$ or $[0, 100]$, but the choice of the interval is not at all a constraint. The end-points are often labeled in accordance with their meaning referring to the degree of satisfaction, quality, agreement, and so on.
2. The *core*, or 1-level set, associated with the response is determined. It corresponds to the interval consisting of the real values within the reference one which are considered to be as 'fully compatible' with the response.
3. The *support*, or its closure or 0-level set, associated with the response is determined. It corresponds to the interval consisting of the real values within the referential that are considered to be as 'compatible to some extent' with the response.
4. The two preceding intervals are 'linearly interpolated' to get a trapezoidal fuzzy number.



In accordance with Likert scales, people respond to items by specifying their feeling with respect to a statement on a symmetric 'agree-disagree', or 'extremely high-extremely low', etc., scale. This specification is performed by choosing one among several given points representing some key degrees of agreement/suitability, etc. To analyze Likert scale-based responses, such points are encoded by means of consecutive integer numbers.

The question posed in Sect. 1, about whether or not fuzzy rating scale-based responses could be into $k$-point Likert's ones, is to be answered in this paper by considering the distance-based mapping $\iota : \mathcal{F}_c^*(\mathbb{R}) \to [a, b]_k = \{a, a + (b - a)/(k - 1), \ldots, a + (b - a)(k - 2)/(k - 1), b\}$ (with $[a, b] =$ reference interval, so that the integer consecutive codes have been re-scaled in accordance with the reference interval) such that $\widetilde{U} \mapsto \arg\min_{i \in [a,b]_k} \rho_2(\widetilde{U}, \mathbb{1}_{\{i\}})$, that is,

$$\iota(\widetilde{U}) = \arg \min_{i \in [a,b]_k} \sqrt{\int_{[0,1]} \frac{(\inf \widetilde{U}_\alpha - i)^2 + (\sup \widetilde{U}_\alpha - i)^2}{2} \, d\alpha},$$

$\rho_2$ being the well-known $L^2$ metric introduced by Diamond and Kloeden [2].

## 3 Real-Life Examples

In this section, we are going to examine three real-life situations in which questionnaires allowing to choose-draw double Likert type-fuzzy rating type responses have been conducted. In each of the examples, we have determined the percentages of coincidences between the expected Likert response (more concretely, the image of the fuzzy rating response through $\iota$ and the assessed Likert response).

*Example 1* By using an online computerized application an experiment has been performed in which people have been asked for their perception of the relative length of different line segments with respect to a pattern longer one (see http://bellman.ciencias.uniovi.es/SMIRE/Perceptions.html).

On the center top of the screen the longest (pattern) line segment has been drawn in black. This segment is fixed for all the trials, so that there is always a reference for the maximum length. At each trial a grey shorter line segment is generated and placed below the pattern one, parallel and without considering a concrete location (i.e., indenting or centering). For each respondent, line segments are generated at random, although to avoid the variation in the perception of different respondents can be mainly due to the variation in length of different generated segments, the (27 first) trials for two respondents refer to the same segments but appearing in different position and location.

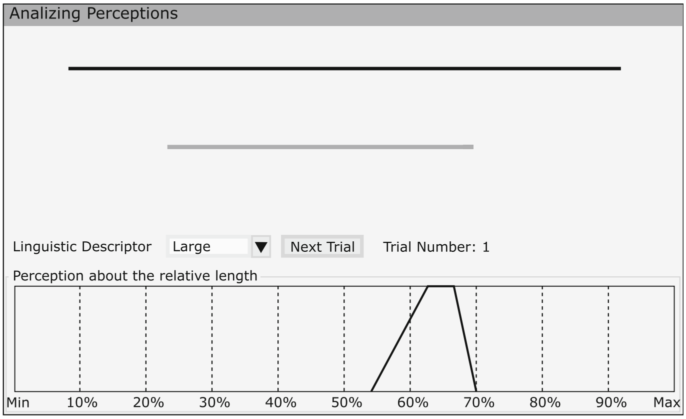
The computerized application explains the formalization and meaning of the fuzzy rating values (see Fig. 1), with reference interval [0, 100]. People have participated online by providing with their judgement of relative length for each of several line segments. Each of these judgements can be doubly expressed: by choosing a label from a 5-point Likert-like list (0 = VERY SMALL, 25 = SMALL, 50 = MEDIUM, 75 = LARGE, 100 = VERY LARGE), and by using the fuzzy rating method.

25 respondents (all with a university scientific background) have supplied 1387 double responses after the corresponding trials. The dataset can be found in http://bellman.ciencias.uniovi.es/smire/Archivos/Perceptionsdataset.pdf. The percentage of coincidences through the minimum distance criterion equals 84.93 %.

*Example 2* A sample of 70 people with different age, background and work type and position has been considered to fill a restaurant customer satisfaction questionnaire with 14 items by using a double response-type form (see http://bellman.ciencias.uniovi.es/smire/FuzzyRatingScaleQuestionnaire-Restaurants.html).

The questionnaire has been conducted by a few students of a Master on Soft Computing and Intelligent Data Analysis held in Mieres in 2011–2012. Figure 2 displays the excerpt of the form to be filled corresponding to one of the involved items.

The form allows the double response, where Likert-like ones are chosen from a 5-point Likert scale (0 = STRONGLY DISAGREE, 25 = SOMEWHAT DISAGREE, 50 = NEUTRAL, 75 = SOMEWHAT AGREE, 100 = STRONGLY AGREE) and the fuzzy ones have reference interval [0, 100].



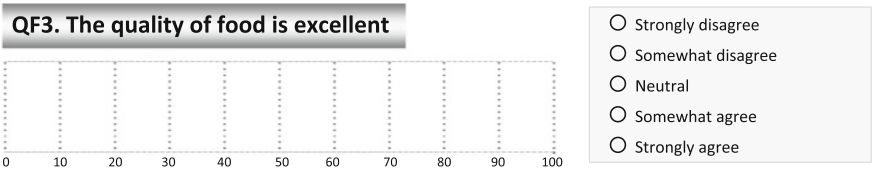
Analizing Perceptions

Linguistic Descriptor: Large

Trial Number: 1

Perception about the relative length

**Fig. 1** Example of a double response from the computerized application in Example 1



QF3. The quality of food is excellent

○ Strongly disagree
○ Somewhat disagree
○ Neutral
○ Somewhat agree
○ Strongly agree

**Fig. 2** Excerpt of a questionnaire about the satisfaction with the quality of restaurants in Example 2

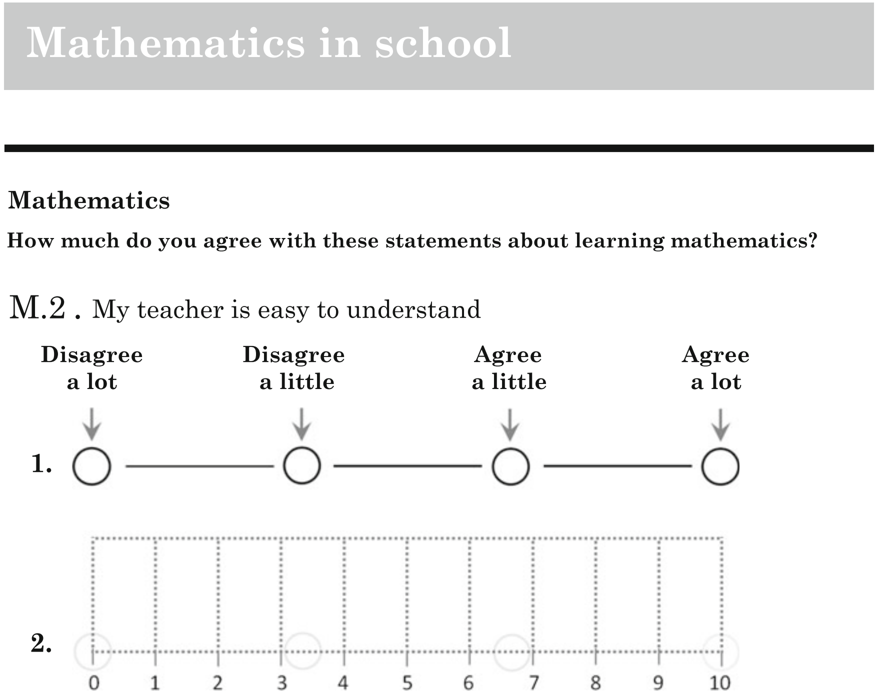
The dataset with 980 double responses can be also found in the webpage including the form. The percentage of coincidences through the minimum distance criterion equals 78.16 %.

*Example 3* This third example is related to the well-known questionnaire TIMSS-PIRLS 2011 which is conducted on populations of (nine to ten years old) fourth grade students and concerns their opinion and feeling on aspects regarding reading, math, and science. This questionnaire is rather standard and most of the involved questions have to be answered according to a 4-point Likert scale ($0 =$ DISAGREE A LOT, $10/3 =$ DISAGREE A LITTLE, $20/3 =$ AGREE A LITTLE, $10 =$ AGREE A LOT).

The original questionnaire form has been adapted to allow a double-type response, the original Likert and a fuzzy rating scale-based one with reference interval $[0, 10]$ (see Fig. 3 for one of the items, and the webpage http://bellman.ciencias.uniovi.es/SMIRE/FuzzyRatingScaleQuestionnaire-SanIgnacio.html for the full paper-and-pencil and computerized forms and datasets).

As a differential feature and to ease the relationship between the two scales for respondents, each numerically encoded Likert response has been superimposed upon the reference interval of the fuzzy rating scale part.

The questionnaire involving these double-response questions has been conducted on 69 fourth grade students from Colegio San Ignacio (Oviedo-Asturias, Spain). The



Mathematics in school

**Mathematics**

**How much do you agree with these statements about learning mathematics?**

M.2 . My teacher is easy to understand

**Disagree a lot** **Disagree a little** **Agree a little** **Agree a lot**

**1.** ○ —— ○ —— ○ —— ○

**2.** 0 1 2 3 4 5 6 7 8 9 10

**Fig. 3** Example of the double-response form to a question in Example 3

dataset with 599 double responses can be also found in the webpage including the form. The percentage of coincidences through the minimum distance criterion equals 81.47 %.

The above indicated percentages for the three examples have been also computed with some other few metrics, even some ones assessing different weights to different $\alpha$-levels (more concretely, assessing weights so that the higher the $\alpha$ the higher/lower the weight). Percentages have been scarcely affected by the choice of the metric.

## 4 Some Remarks from the Analysis of the Real-Life Examples

As a summary of the analysis of the percentages in the three examples in Sect. 3 we can empirically conclude that background, age and sample sizes seem not to be very influential, as we could formerly suspect. Actually, we should confess that children in the third example, which are much younger and are assumed not to have yet a high background, have positively surprised us with their ease to catch the idea in just 15 min of explanation.

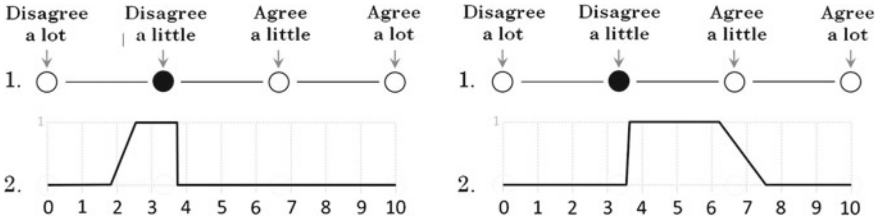


**Fig. 4** Example of two fuzzy responses from Example 3 for which both the real and the minimum distance-based Likert labels coincide

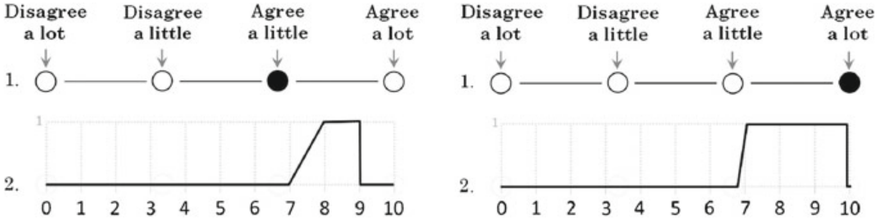


**Fig. 5** Example of two fuzzy responses from Example 3 for which the minimum distance-based Likert labels coincide, but the real choices do not

On the other hand, we can also conclude that in real-life people having the opportunity of the double response is not necessarily guided by what Likert labels can mean. In fact, it seems that people take advantage of the flexibility, freedom and expressiveness of the fuzzy rating scale to draw their valuations and they make it rather independently of their Likert assessment even in case they have to do it simultaneously. This corroborates what has been statistically concluded by Lubiano et al. [6, 7]: Likert scales 'aggregate' in some sense valuations which could be 'precisiated' through fuzzy numbers, so relevant information can be lost when using Likert scales.

This paper also adds that the real-life aggregation does not correspond in practice to a natural (distance-based) partition of the fuzzy rating scale-based responses. And, probably, there is no criterion which could properly mimic human association. In this way, the following responses have been taken from the dataset of the responses in Example 3 to the item $M.2$ in Fig. 3, namely, "My math teacher is easy to understand". Figure 4 shows two very different fuzzy responses to this item for which both the distance-based association and the real choice from the 4-point Likert scale coincide (DISAGREE A LITTLE). Figure 5 shows two rather close fuzzy responses to this item for which the distance-based association from the 4-point Likert scale coincide (AGREE A LITTLE), but the real choices do not.

To end this paper, we would like simply illustrating these conclusions with a simple instance also taken from the dataset of the responses in Example 3 to the item $M.2$ in Fig. 3. Among the 69 double responses to this item, 10 of the Likert components have not matched with the minimum distance Likert (that we can refer

**Table 1** Responses to the item "My math teacher is easy to understand" in Example 3 for which the real 4-point Likert choice and the minimum distance one do not match

| inf $\widetilde{U}_0$ | inf $\widetilde{U}_1$ | sup $\widetilde{U}_1$ | sup $\widetilde{U}_0$ | Chosen Likert | dist to 0 | dist to 10/3 | dist to 20/3 | dist to 10 | Mindist Likert $\iota$ | Width support | Width core |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 3.55 | 6.25 | 7.5 | 10/3 | 5.47 | 2.52 | 2.24 | 5.09 | 20/3 | **4** | 2.7 |
| 5.95 | 6 | 9.2 | 10 | 10 | 8 | 4.81 | 2.14 | 2.86 | 20/3 | **4.05** | 3.2 |
| 4.9 | 4.9 | 8.45 | 9.975 | 10 | 7.38 | 4.31 | 2.21 | 3.66 | 20/3 | **5.075** | 3.55 |
| 8 | 8.5 | 8.5 | 9 | 20/3 | 8.5 | 5.17 | **1.86** | **1.53** | 10 | 1 | 0 |
| 3.4 | 4.825 | 9.95 | 9.95 | 10 | 7.62 | 4.72 | 2.96 | 4.17 | 20/3 | **6.55** | 5.125 |
| 3.175 | 5.025 | 7.5 | 9.95 | 10 | 6.85 | 3.9 | 2.41 | 4.31 | 20/3 | **6.775** | 2.475 |
| 8 | 8.5 | 9.2 | 9.2 | 20/3 | 8.74 | 5.41 | **2.11** | **1.36** | 10 | 1.2 | 0.7 |
| 5.6 | 6.7 | 9.15 | 10 | 10 | 8.05 | 4.85 | 2.11 | 2.75 | 20/3 | **4.4** | 2.45 |
| 5.825 | 5.85 | 9.875 | 9.95 | 10 | 8.13 | 4.98 | 2.37 | 2.94 | 20/3 | **4.125** | 4.025 |
| 2.5 | 4.625 | 4.625 | 6.9 | 20/3 | 4.83 | 1.84 | 2.37 | 5.49 | 10/3 | **4.4** | 0 |

to as the expected Likert label). These responses have been gathered in Table 1, where we can **easily see** that 8 of them correspond to the 8 widest (w.r.t. support, and, mostly, w.r.t. core) fuzzy responses, whereas the other 2 correspond to narrower fuzzy responses but showing close distances (w.r.t. the maximum distance 10) to two of the encoded Likert responses.

Finally, it should be emphasized that the high percentage of coincidences of the real and the minimum distance-based 'Likertization' processes should not be viewed as an argument in favour of the use of the Likert scale in contrast to the fuzzy rating one. On the contrary, situations like those in Figs. 4 and 5 clearly illustrate the need for the last scale, whenever it can be properly employed and data are to be statistically analyzed. Thus, the statistical analysis of the Likert responses in Fig. 4 doesn't distinguish between them, whereas the responses are indisputably different if the fuzzy rating scale is considered. Consequently, many errors, deviations, differences, are often neglected in using Likert scales.

**Acknowledgments** Authors are grateful to Colegio San Ignacio in Oviedo-Asturias (Spain) for allowing us to collect the data in the real-life example. The research in this paper has been partially supported by/benefited from Principality of Asturias Grants GRUPIN14-101 and Severo Ochoa BP12012 (De la Rosa de Sáa), and the Spanish Ministry of Economy and Competitiveness Grants MTM2015-63971-P and MTM2013-44212-P. Their financial support is gratefully acknowledged.

## References

1. De la Rosa de Sáa S, Gil MA, González-Rodríguez G, López MT, Lubiano MA (2015) Fuzzy rating scale-based questionnaires and their statistical analysis. IEEE T Fuzzy Syst 23:111–126
2. Diamond P, Kloeden P (1999) Metric spaces of fuzzy sets. Fuzzy Sets Syst 100:63–71
3. Gil MA, Lubiano MA, De la Rosa de Sáa S, Sinova B (2015) Analyzing data from a fuzzy rating scale-based questionnaire. A case study. Psicothema 27:182–191

4. Hesketh T, Hesketh B (1994) Computerized fuzzy ratings: the concept of a fuzzy class. Behav Res Meth Ins C 26:272–277
5. Hesketh T, Pryor R, Hesketh B (1988) An application of a computerized fuzzy graphic rating scale to the psychological measurement of individual differences. Int J Man-Mach Stud 29:21–35
6. Lubiano MA, De la Rosa de Sáa S, Montenegro M, Sinova B, Gil MA (2016) Descriptive analysis of responses to items in questionnaires. Why not using a fuzzy rating scale? Inform Sci 360:131–148
7. Lubiano MA, Montenegro M, Sinova B, De la Rosa de Sáa S, Gil MA (2016) Hypothesis testing for means in connection with fuzzy rating scale-based data: algorithms and applications. Eur J Oper Res 251:918–929

# Asymptotic Results for Sums of Independent Random Variables with Alternating Laws

**Claudio Macci**

**Abstract** Stochastic models governed by alternating dynamics arise in various applications. In several cases these models can be described by sums of independent random variables with alternating laws. The aim of this paper is to study the asymptotic behavior of these sums in the fashion of large deviations.

## 1 Introduction

Stochastic models with alternating dynamics arise in various applications and are widely studied in the literature. A remarkable example is telegraph process (see e.g. [5]; see also [1] for the case with drift) which is considered in several fields; for instance see [6] for its use in finance modeling (a wide source for other similar models in this field is [4]).

In several cases these models are well described by sequences of sums $\left\{\sum_{i=1}^{n} X_i : n \geq 1\right\}$ of independent random variables $\{X_n : n \geq 1\}$ with alternating laws; namely we mean that the odd summands $\{X_{2n-1} : n \geq 1\}$ are identically distributed with law $\xi$, and the even summands $\{X_{2n} : n \geq 1\}$ are identically distributed with law $\nu$. This kind of sequences has been recently studied in [7] where $\xi$ and $\nu$ are exponential laws, i.e.

$$\xi(dx) = \lambda_\xi e^{-\lambda_\xi x} 1_{(0,\infty)}(x)dx \quad \text{and} \quad \nu(dx) = \lambda_\nu e^{-\lambda_\nu x} 1_{(0,\infty)}(x)dx$$

for some $\lambda_\xi, \lambda_\nu > 0$.

The aim of this paper is to prove large deviation results for these sums. The theory of large deviations gives an asymptotic computation of small probabilities on exponential scale (see [3] as a reference on this topic). The results are presented in Sect. 3 after some preliminaries in Sect. 2. The final Sect. 4 is devoted to some concluding remarks.

C. Macci (✉)
Dipartimento di Matematica, Università di Roma Tor Vergata, Rome, Italy
e-mail: macci@mat.uniroma2.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_42

## 2 Preliminaries

We start by recalling some basic definitions on large deviations. A speed function is a sequence $\{v_n : n \geq 1\}$ such that $\lim_{n\to\infty} v_n = \infty$. Let $\mathcal{Z}$ be a Hausdorff topological space with Borel $\sigma$-algebra $\mathcal{B}(\mathcal{Z})$ (here we always assume that $\mathcal{Z} = \mathbb{R}$); a lower semicontinuous function $I : \mathcal{Z} \to [0, \infty]$ is called rate function. A sequence of $\mathcal{Z}$-valued random variables $\{Z_n : n \geq 1\}$ (defined on the same probability space $(\Omega, \mathcal{F}, P)$) satisfies the *large deviation principle* (LDP for short) with speed $v_n$ and rate function $I$ if

$$\limsup_{n\to\infty} \frac{1}{v_n} \log P(Z_n \in C) \leq -\inf_{z\in C} I(z) \quad \text{for all closed sets } C \subset \mathcal{Z}$$

and

$$\liminf_{n\to\infty} \frac{1}{v_n} \log P(Z_n \in G) \geq -\inf_{z\in G} I(z) \quad \text{for all open sets } G \subset \mathcal{Z}.$$

We remark that the definition of LDP concerns the laws of the random variables $\{Z_n : n \geq 1\}$; therefore the random variables $\{Z_n : n \geq 1\}$ do not need to be defined on the same probability space. Finally a rate function $I$ is said to be good if all its level sets $\{\{z \in \mathcal{Z} : I(z) \leq \eta\} : \eta \geq 0\}$ are compact.

The term *moderate deviations* is used for a class of LDPs determined by the sequences of positive numbers $\{a_n : n \geq 1\}$ such that Eq. (3) below holds (see Theorem 3); these LDPs concern centered random variables and are governed by the same quadratic rate function which vanishes at the origin only. In some sense moderate deviations fill the gap between a law of large numbers for centered random variables, and an asymptotic Normality result; this aspect will be illustrated in Sect. 4 (see Remark 2).

In view of what follows we recall that a convex function $f : \mathbb{R} \to (-\infty, \infty]$ is *essentially smooth* if:

- the interior $\mathcal{D}_f^\circ$ of its domain $\mathcal{D}_f := \{\theta \in \mathbb{R} : f(\theta) < \infty\}$ is nonempty;
- the function $f$ is differentiable throughout $\mathcal{D}_f^\circ$;
- the function $f$ is steep (namely, if $|f'(\theta_n)| \to \infty$ as $n \to \infty$ whenever $\{\theta_n : n \geq 1\} \subset \mathcal{D}_f^\circ$ approaches to the boundary of $\mathcal{D}_f$ as $n \to \infty$).

We remark that the steepness condition holds vacuously if the function $f$ is finite and differentiable everywhere.

Now we are ready to recall the statement of the well-known Gärtner Ellis Theorem (see e.g. Theorem 2.3.6(c) in [3]) and, for our aim, we restrict the attention on the case of real valued random variables.

**Theorem 1** (Gärtner Ellis Theorem) *Let $\{Z_n : n \geq 1\}$ be a sequence of real valued random variables such that the limit*

$$\Lambda(\theta) := \lim_{n\to\infty} \frac{1}{v_n} \log \mathbb{E}\left[e^{v_n \theta Z_n}\right]$$

*exists as an extended real number for all $\theta \in \mathbb{R}$. Moreover, assume that $0 \in \mathcal{D}_{\Lambda}^{\circ}$. Then, if the function $\Lambda$ is essentially smooth and lower semi-continuous, the sequence $\{Z_n : n \geq 1\}$ satisfies the LDP with speed $v_n$ and good rate function $\Lambda^*$ defined by $\Lambda^*(x) := \sup_{\theta \in \mathbb{R}}\{x\theta - \Lambda(\theta)\}$.*

## 3 Large and Moderate Deviations

Here we always consider probability measures $\pi$ on $\mathbb{R}$ such that

$$\Lambda_\pi(\theta) := \log \int_{\mathbb{R}} e^{\theta x} \pi(dx)$$

is finite in a neighborhood of $\theta = 0$, essentially smooth and we have

$$\Lambda_\pi(\theta) = \theta \Lambda_\pi'(0) + \frac{\theta^2}{2} \Lambda_\pi''(0) + o(\theta^2), \tag{1}$$

where $\frac{o(\theta^2)}{\theta^2} \to 0$ as $\theta \to 0$. It is known that, in this case, $\Lambda_\pi'(0)$ and $\Lambda_\pi''(0)$ are mean and variance of any random variable with law $\pi$; thus

$$\Lambda_\pi''(0) \geq 0. \tag{2}$$

We recall the logarithm of a moment generating function is always a lower semi-continuous (see e.g. Exercise 2.3.16(a) in [3]). Moreover, we set

$$\Lambda_\pi^*(x) := \sup_{\theta \in \mathbb{R}}\{x\theta - \Lambda_\pi(\theta)\}.$$

In our results we always assume that:

- $\xi$ and $\nu$ are two different probability measures on $\mathbb{R}$ which satisfy the above hypotheses presented for $\pi$.
- $\{X_n : n \geq 1\}$ are independent real valued random variables such that $\{X_{2n-1} : n \geq 1\}$ are identically distributed with law $\xi$, and $\{X_{2n} : n \geq 1\}$ are identically distributed with law $\nu$.

We remark that we assume that $\xi \neq \nu$ to avoid a well-known case (see Remark 3 in Sect. 4); furthermore, with some slight changes of the proofs, Theorems 2 and 3 still hold if the roles of $\xi$ and $\nu$ are exchanged (i.e. if $\{X_{2n-1} : n \geq 1\}$ are identically distributed with law $\nu$, and $\{X_{2n} : n \geq 1\}$ are identically distributed with law $\xi$).

We recall that, for the convolution $\xi \diamond \nu$ between $\xi$ and $\nu$, we have $\Lambda_{\xi \diamond \nu} = \Lambda_\xi + \Lambda_\nu$ and, under our hypotheses, $\xi \diamond \nu$ satisfies the hypotheses presented above for $\pi$.

**Theorem 2** (Large Deviations) *Let* $\{X_n : n \geq 1\}$ *be a sequence of real valued random variables as above. Then the sequence* $\left\{\frac{1}{n}\sum_{i=1}^n X_i : n \geq 1\right\}$ *satisfies the LDP with speed* $v_n = n$ *and good rate function* $J$ *defined by* $J(x) := \frac{1}{2}\Lambda^*_{\xi\diamond\nu}(2x)$.

*Proof* We want to apply Theorem 1 with $v_n = n$ and $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$. For all $\theta \in \mathbb{R}$ we distinguish two cases: if $n$ is even we have

$$\frac{1}{n}\log \mathbb{E}\left[e^{n\theta\frac{1}{n}\sum_{i=1}^n X_i}\right] = \frac{1}{n}\cdot\frac{n}{2}\Lambda_{\xi\diamond\nu}(\theta)$$

whereas, if $n$ is odd, we have

$$\frac{1}{n}\log \mathbb{E}\left[e^{n\theta\frac{1}{n}\sum_{i=1}^n X_i}\right] = \frac{1}{n}\left(\frac{n-1}{2}\Lambda_{\xi\diamond\nu}(\theta) + \Lambda_\xi(\theta)\right).$$

Then, since

$$\frac{1}{n}\Lambda_\xi(\theta) \to \begin{cases} 0 & \text{if } \Lambda_\xi(\theta) < \infty \\ \infty & \text{if } \Lambda_\xi(\theta) = \infty, \end{cases}$$

and $\{\theta \in \mathbb{R} : \Lambda_{\xi\diamond\nu}(\theta) < \infty\} \subset \{\theta \in \mathbb{R} : \Lambda_\xi(\theta) < \infty\}$ (because $\Lambda_{\xi\diamond\nu} = \Lambda_\xi + \Lambda_\nu$), for all $\theta \in \mathbb{R}$ we have

$$\lim_{n\to\infty}\frac{1}{n}\log \mathbb{E}\left[e^{n\theta\frac{1}{n}\sum_{i=1}^n X_i}\right] = \frac{1}{2}\Lambda_{\xi\diamond\nu}(\theta).$$

In conclusion, by Theorem 1, $\left\{\frac{1}{n}\sum_{i=1}^n X_i : n \geq 1\right\}$ satisfies the LDP with speed $v_n = n$ and good rate function $J$ defined by

$$J(x) := \sup_{\theta\in\mathbb{R}}\left\{x\theta - \frac{1}{2}\Lambda_{\xi\diamond\nu}(\theta)\right\},$$

and one can easily see that it coincides with the rate function in the statement of proposition. □

**Theorem 3** (Moderate Deviations) *Let* $\{X_n : n \geq 1\}$ *be a sequence of real valued random variables as above. Then, for all sequences of positive numbers* $\{a_n : n \geq 1\}$ *such that*

$$\lim_{n\to\infty} a_n = 0 \textit{ and } \lim_{n\to\infty} na_n = \infty, \tag{3}$$

*the sequence* $\left\{\sqrt{na_n}\cdot\frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{n} : n \geq 1\right\}$ *satisfies the LDP with speed* $v_n = 1/a_n$ *and good rate function* $\tilde{J}$ *defined by* $\tilde{J}(x) := \frac{x^2}{\Lambda''_{\xi\diamond\nu}(0)}$ *if* $\Lambda''_{\xi\diamond\nu}(0) > 0$, *and by*

$$\tilde{J}(x) := \begin{cases} 0 & \textit{if } x = 0 \\ \infty & \textit{if } x \neq 0 \end{cases}$$

*if* $\Lambda''_{\xi\diamond\nu}(0) = 0$.

*Proof* We want to apply Theorem 1 with $v_n = 1/a_n$ and $Z_n = \sqrt{na_n} \cdot \frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{n}$. For all $\theta \in \mathbb{R}$ we have to consider

$$\Psi_n(\theta) := \frac{1}{1/a_n} \log \mathbb{E}\left[e^{\frac{\theta}{a_n} \cdot \sqrt{na_n} \cdot \frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{n}}\right],$$

and therefore (after simple computations)

$$\Psi_n(\theta) = a_n \left( \log \mathbb{E}\left[e^{\theta \cdot \frac{\sum_{i=1}^n X_i}{\sqrt{na_n}}}\right] - \frac{\theta}{\sqrt{na_n}} \sum_{i=1}^n \mathbb{E}[X_i] \right).$$

We remark that, for all $\theta \in \mathbb{R}$, $\frac{\theta}{\sqrt{na_n}} \to 0$ as $n \to \infty$; therefore, for $n$ large enough, $\Lambda_{\xi\diamond\nu}\left(\frac{\theta}{\sqrt{na_n}}\right)$ and $\Lambda_\xi\left(\frac{\theta}{\sqrt{na_n}}\right)$ are finite and we can consider Eq. (1) with $\pi = \xi \diamond \nu$ and with $\pi = \xi$. We distinguish two cases. If $n$ is even we have

$$\Psi_n(\theta) = \frac{a_n n}{2} \left( \Lambda_{\xi\diamond\nu}\left(\frac{\theta}{\sqrt{na_n}}\right) - \frac{\theta}{\sqrt{na_n}} \Lambda'_{\xi\diamond\nu}(0) \right)$$

and therefore

$$\Psi_n(\theta) = \frac{a_n n}{2} \left( \frac{1}{2} \cdot \frac{\theta^2}{na_n} \Lambda''_{\xi\diamond\nu}(0) + o\left(\frac{\theta^2}{na_n}\right) \right).$$

If $n$ is odd we have

$$\begin{aligned} \Psi_n(\theta) &= a_n \left( \frac{n-1}{2} \Lambda_{\xi\diamond\nu}\left(\frac{\theta}{\sqrt{na_n}}\right) + \Lambda_\xi\left(\frac{\theta}{\sqrt{na_n}}\right) \right) \\ &\quad - a_n \left( \frac{\theta}{\sqrt{na_n}} \left( \frac{n-1}{2} \Lambda'_{\xi\diamond\nu}(0) + \Lambda'_\xi(0) \right) \right) \\ &= \frac{a_n(n-1)}{2} \left( \Lambda_{\xi\diamond\nu}\left(\frac{\theta}{\sqrt{na_n}}\right) - \frac{\theta}{\sqrt{na_n}} \Lambda'_{\xi\diamond\nu}(0) \right) \\ &\quad + a_n \left( \Lambda_\xi\left(\frac{\theta}{\sqrt{na_n}}\right) - \frac{\theta}{\sqrt{na_n}} \Lambda'_\xi(0) \right) \\ &= \frac{a_n(n-1)}{2} \left( \frac{1}{2} \cdot \frac{\theta^2}{na_n} \Lambda''_{\xi\diamond\nu}(0) + o\left(\frac{\theta^2}{na_n}\right) \right) \\ &\quad + a_n \left( \Lambda_\xi\left(\frac{\theta}{\sqrt{na_n}}\right) - \frac{\theta}{\sqrt{na_n}} \Lambda'_\xi(0) \right). \end{aligned}$$

Then we can say that

$$\lim_{n\to\infty} \Psi_n(\theta) = \frac{\theta^2}{4}\Lambda''_{\xi\diamond\nu}(0). \tag{4}$$

In conclusion, by Theorem 1, $\left\{\sqrt{na_n}\cdot\frac{\sum_{i=1}^n(X_i-\mathbb{E}[X_i])}{n} : n \geq 1\right\}$ satisfies the LDP with speed $v_n = 1/a_n$ and good rate function $\tilde{J}$ defined by

$$\tilde{J}(x) := \sup_{\theta\in\mathbb{R}}\left\{x\theta - \frac{\theta^2}{4}\Lambda''_{\xi\diamond\nu}(0)\right\}.$$

Furthermore, $\Lambda''_{\xi\diamond\nu}(0) \geq 0$ by Eq. (2) (with $\pi = \xi\diamond\nu$); on the other hand we have $\Lambda''_{\xi\diamond\nu}(0) \geq 0$ because the function $\theta \mapsto \frac{\theta^2}{4}\Lambda''_{\xi\diamond\nu}(0)$ is a convex (in fact we have a limit of convex functions in Eq. (4)). Then, if we distinguish the cases $\Lambda''_{\xi\diamond\nu}(0) > 0$ and $\Lambda''_{\xi\diamond\nu}(0) = 0$, we easily get the expressions of $\tilde{J}$ in the statement of the proposition.

□

## 4 Concluding Remarks

This section is devoted to some concluding remarks on Theorems 2 and 3, and on the case $\xi = \nu$.

*Remark 1* (*On Theorem* 2) It is known (and this easily can be checked) that $\Lambda^*_{\xi\diamond\nu}(x) = 0$ if and only if $x = \Lambda'_{\xi\diamond\nu}(0) = \Lambda'_\xi(0) + \Lambda'_\nu(0)$. Therefore, as far as the rate function $J$ in Theorem 2 is concerned, we have $J(x) = 0$ if and only if $x = x_\infty$, where

$$x_\infty := \frac{1}{2}\left(\Lambda'_\xi(0) + \Lambda'_\nu(0)\right)$$

is the mean of the expected values of two random variables with laws $\xi$ and $\nu$. The LDP in Theorem 2 allows to say that the sequence $\left\{\frac{1}{n}\sum_{i=1}^n X_i : n \geq 1\right\}$ converges to $x_\infty$ (as $n \to \infty$). In fact, for all open sets $A$ such that $x_\infty \in A$, we have $J(A^c) := \inf_{x\in A^c} J(x) > 0$ and, for all $\varepsilon > 0$, there exists $n_\varepsilon$ such that

$$P\left(\frac{1}{n}\sum_{i=1}^n X_i \in A^c\right) \leq e^{-n(J(A^c)-\varepsilon)}$$

for all $n > n_\varepsilon$.

*Remark 2* (*On Theorem* 3) Firstly we can say that $\Lambda''_{\xi\diamond\nu}(0) = \Lambda''_\xi(0) + \Lambda''_\nu(0)$ is nonnegative because is the sum of two variances; so we can have $\Lambda''_{\xi\diamond\nu}(0) = 0$ if and only if both $\xi$ and $\nu$ are degenerating probability measures (i.e. the laws of constant random variables). On the contrary, if $\Lambda''_{\xi\diamond\nu}(0) > 0$ because at least one variance is strictly positive, we have

$$\tilde{J}(x) := \frac{x^2}{2\sigma^2},$$

where $\sigma^2 := \frac{1}{2}(\Lambda''_\xi(0) + \Lambda''_\nu(0))$ it is the mean of the variances of two random variables with laws $\xi$ and $\nu$.

Moreover, a closer inspection of the proof of Theorem 3 reveals that the relation in Eq. (4) holds even if $a_n = 1$ for all $n \geq 1$, i.e.

$$\lim_{n\to\infty} \log \mathbb{E}\left[e^{\theta\cdot\frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{n}}}\right] = \frac{\theta^2}{4}\Lambda''_{\xi\diamond\nu}(0) = \frac{\theta^2}{2}\sigma^2;$$

therefore we can say that $\frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{n}}$ converges weakly to the centered Normal distribution with variance $\sigma^2$ (as $n \to \infty$).

Thus, in some sense, moderate deviations fill the gap between this weak convergence to the centered Normal distribution with variance $\sigma^2$, and the convergence of $\frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{n}$ to 0 (as $n \to \infty$). In the first case, as we said above, we have $a_n = 1$ and in the second case we have $a_n = \frac{1}{n}$ (for all $n \geq 1$); thus, in both cases, one condition in Eq. (3) holds and the other one fails.

In connection with the arguments of this remark we recall the paper [2] where an asymptotic Normality result can be derived by a LDP proved by using Gärtner Ellis Theorem (i.e. Theorem 1 in this paper).

*Remark 3* (*On the case* $\xi = \nu$) In this case the random variables $\{X_n : n \geq 1\}$ are identically distributed with law $\xi = \nu$. Thus, if we look at the proof of Theorem 2 in this paper, we have

$$\frac{1}{2}\Lambda_{\xi\diamond\nu} = \frac{1}{2}(\Lambda_\xi + \Lambda_\nu) = \Lambda_\xi,$$

or $\frac{1}{2}\Lambda_{\xi\diamond\nu} = \Lambda_\nu$, and therefore the LDP holds with good rate function $J = \Lambda^*_\xi = \Lambda^*_\nu$. We also remark that, when we deal to the i.i.d. case, we can directly refer to Theorems 2.2.3 and 3.7.1 in [3] (for large and moderate deviations, respectively). Moreover, Theorem 2.2.3 in [3] (i.e. the well-known Cramér Theorem) provides the LDP with rate function $J = \Lambda^*_\xi = \Lambda^*_\nu$ even without having steep logarithm moment generating functions and the goodness of the rate function could fail.

## References

1. Beghin L, Nieddu L, Orsingher E (2001) Probabilistic analysis of the telegrapher's process with drift by means of relativistic transformations. J Appl Math Stochastic Anal 14:11–25
2. Bryc W (1993) A remark on the connection between the large deviation principle and the central limit theorem. Stat Probab Lett 18:253–256

3. Dembo A, Zeitouni O (1998) Large deviations techniques and applications, 2nd edn. Springer, New York
4. Kolesnik AD, Ratanov N (2013) Telegraph processes and option pricing. Springer, Heidelberg
5. Orsingher E (1990) Probability law, flow function, maximum distribution of wave-governed random motions and their connections with Kirchoff's laws. Stochastic Process Appl 34:49–66
6. Ratanov N (2007) A jump telegraph model for option pricing. Quant Finance 7:575–583
7. Ratanov N (2015) Hypo-exponential distributions and compound Poisson processes with alternating parameters. Stat Probab Lett 107:71–78

# Dispersion Measures and Multidistances on $\mathbb{R}^k$

**Javier Martín and Gaspar Mayor**

**Abstract** After introducing a definition of dispersion measure on the Euclidean space $\mathbb{R}^k$, we deal with the connection between these measures and the so called multidistances. In this way, we show that thr standard deviation is a relevant example of multidistance and, on the other hand, several significant families of multidistances are, at the same time, dispersion measures. Sufficient conditions for a multidistance to be a dispersion measure are also established.

## 1 Introduction

Descriptive Statistics provides some indexes to measure the dispersion of a set of unidimensional data. Several attempts have been done in order to set a general framework to deal with this topic, introducing different axiomatic definitions, such as [5].

According to the fact that in many situations the data to be processed are multidimensional in nature, it seems reasonable to have a tool which also allows measuring the dispersion of a set of such data. In this contribution we introduce an axiomatic definition of dispersion measures, based on the one given in [2], and we study the relationship between these measures and the multi-argument distances, multidistances for short, defined in [3].

The paper is organized as follows. Section 2 introduces our proposal of axiomatic definition of dispersion measure, and compares it with the one given in [2]. In Sect. 3 we recall the definition of multidistance and prove that the standard deviation is a relevant example of this kind of multidimensional distances. Then, we prove that functionally expressible multidistances, fulfilling an additional condition, are dispersion measures. Finally, multidistances belonging to three relevant families are shown to be also dispersion measures.

J. Martín (✉) · G. Mayor
University of the Balearic Islands, Palma, Spain
e-mail: javier.martin@uib.es

G. Mayor
e-mail: gmayor@uib.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_43

## 2 Measures of Dispersion on $\mathbb{R}^k$

Let us consider in this work the set $\left(\mathbb{R}^k\right)^n$ of all finite lists of elements of $\mathbb{R}^k$.

We recall here the definition of measure of dispersion given in [2].

**Definition 1** A function $\varDelta\colon \bigcup_{n\geqslant 1}\left(\mathbb{R}^k\right)^n \to \mathbb{R}^+$ is called a measure of dispersion if $\varDelta$ is not identically zero and it satisfies the following axioms for all $n \geqslant 1$ and all $(x_1, \ldots, x_n)$ in $\left(\mathbb{R}^k\right)^n$:

(A1) $\varDelta(x, \ldots, x) = 0$.
(A2) $\varDelta$ is symmetric.
(A3) $\varDelta$ is invariant under translations.
(A4) $\varDelta$ is invariant under rotations.

Sometimes, the authors point out, one more axiom is also considered:

(A5) There exists a function $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ such that, for all $a \in \mathbb{R}^+$:

$$\varDelta(ax_1, \ldots, ax_n) = \rho(a)\varDelta(x_1, \ldots, x_n)\ . \tag{1}$$

The definition we propose in this work shares several axioms, and the other ones are slightly modified.

**Definition 2** A function $\varDelta\colon \bigcup_{n\geqslant 1}\left(\mathbb{R}^k\right)^n \to \mathbb{R}^+$ is said to be a $c$-dispersion measure, $c > 0$, when it fulfills these four conditions:

($\varDelta$1) $\varDelta(x_1, \ldots, x_n) = 0$ if and only if $x_i = x_j$ for all $i, j = 1, \ldots, n$.
($\varDelta$2) $\varDelta$ is symmetric: $\varDelta(x_1, \ldots, x_n) = \varDelta(x_{\pi(1)}, \ldots, x_{\pi(n)})$ for any permutation $\pi$ of $\{1, \ldots, n\}$.
($\varDelta$3) $\varDelta$ is invariant under isometries: $\varDelta(x_1, \ldots, x_n) = \varDelta(\phi(x_1), \ldots, \phi(x_n))$ for any isometry $\phi$ of $\mathbb{R}^k$.
($\varDelta$4) $\varDelta(ax_1, \ldots, ax_n) = a^c\varDelta(x_1, \ldots, x_n)$, for all $a \in \mathbb{R}^+$.

*Remark 1* Our definition is, of course, more restrictive than the given by axioms A1 to A5. However, it is worth noting that if we add some very weak hypothesis, such as the continuity at a point of the function $\rho$ in axiom A5, and taking into account the rest of axioms and the condition $\varDelta \neq 0$, it can be deduced that $\rho$ has to be of the form $\rho(a) = a^c$ for all $a > 0$, with $c$ arbitrary. We have only considered positive values of $c$ in condition $\varDelta$4 because of the nature of the concept that we are defining.

*Example 1* The function $\varDelta\colon \bigcup_{n\geqslant 1}\left(\mathbb{R}^k\right)^n \to \mathbb{R}^+$, defined for all list $x_1, \ldots, x_n)$ of elements of $\mathbb{R}^k$ by the formula

$$\varDelta(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\ , \tag{2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and its square root $\sqrt{\Delta}$, are examples of dispersion measures in $\mathbb{R}^k$. They fulfill conditions $\Delta 1$, $\Delta 2$, $\Delta 3$, and taking into account that

$$\Delta^2(ax_1, \ldots, ax_n) = a^2\Delta^2(x_1, \ldots, x_n) \ , \tag{3}$$

condition $\Delta 4$ also holds: $\Delta$ is a 2-dispersion measure and $\sqrt{\Delta}$ is a 1-dispersion measure.

Observe that $\Delta$ and $\sqrt{\Delta}$ generalize the usual variance $\sigma^2$ and standard deviation $\sigma$, respectively, because they are obtained in the case $k = 1$.

## 3 Multidistances and Dispersion Measures

Multidistances are a generalization of ordinary distances in order to measure how much separated are not only two elements of a set but any finite list. They are defined as follows.

**Definition 3** [3] We say that a function $D\colon \bigcup_{n\geqslant 1} X^n \to \mathbb{R}^+$ is a multidistance on a non empty set $X$ when the following properties hold, for all $n$ and $x_1, \ldots, x_n, y \in X$:

(md1) $D(x_1, \ldots, x_n) = 0$ if and only if $x_i = x_j$ for all $i, j = 1, \ldots n$.
(md2) $D(x_1, \ldots, x_n) = D(x_{\pi(1)}, \ldots, x_{\pi(n)})$ for any permutation $\pi$ of $\{1, \ldots, n\}$.
(md3) $D(x_1, \ldots, x_n) \leqslant D(x_1, y) + \cdots + D(x_n, y)$, for all $y \in X$.

A remarkable example is the so called Fermat multidistance:

$$D_F(x_1, \ldots, x_n) = \min\left\{\sum_{i=1}^{n} d(x_i, x)\colon x \in \mathbb{R}^k\right\}, \ \ \forall x_1, \ldots, x_n \in \mathbb{R}^k \ . \tag{4}$$

We will deal with this multidistance at the end of this section.

The next result shows that multidistances and dispersion measures are interrelated notions.

**Proposition 1** *The standard deviation* $\sigma\colon \bigcup_{n\geqslant 1} \mathbb{R}^n \to \mathbb{R}^+$,

$$\sigma(x_1, \ldots, x_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \ , \tag{5}$$

*is a multidistance.*

*Proof* Conditions md1, md2 are trivially fulfilled.

Let us go with condition md3.

- The cases $n = 1, 2$ are also trivial. Observe that for $n = 2$ we have $\sigma(x_1, x_2) = \frac{1}{2}|x_1 - x_2|$.
- For $n \geqslant 4$, we have:

$$\begin{aligned}\sigma(x_1, \ldots, x_n) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &\leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y)^2} \;\; \forall y \in \mathbb{R} \\ &\leqslant \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|x_i - y|^2 \;\; \forall y \in \mathbb{R} \\ &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\sigma(x_i, y) \;\; \forall y \in \mathbb{R} ,\end{aligned}$$

where the first inequality is due to the fact that the mean minimizes the sum of square deviations, the second one is a property of the sum of squares of non negative real numbers, namely $\sum a_i^2 \leqslant \left(\sum a_i\right)^2$, and finally the expression of standard deviation for two numbers has been used in the last equality.
But for all $n \geqslant 4$,

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\sigma(x_i, y) \leqslant \sum_{i=1}^{n}\sigma(x_i, y) \;\; \forall y \in \mathbb{R} ,$$

and so this case is proved.
- For $n = 3$ we have to proof that for all $x_1, x_2, x_3, y \in \mathbb{R}$,

$$\sigma(x_1, x_2, x_3) \leqslant \sigma(x_1, y) + \sigma(x_2, y) + \sigma(x_3, y) \; . \tag{6}$$

Without loss of generality, we can consider that $x_1 \leqslant x_2 \leqslant x_3$, with $x_1 < x_3$.

The transformation defined by $f(t) = \frac{2}{x_3 - x_1}t - \frac{x_3 + x_1}{x_3 - x_1}$ converts the list $(x_1, x_2, x_3)$ into $(-1, \alpha, 1)$, where $\alpha = \frac{2x_2 - x_1 + x_3}{x_3 - x_1} \in [-1, 1]$.

Then,

$$\sigma(x_1, x_2, x_3) = \sigma(f^{-1}(-1), f^{-1}(\alpha), f^{-1}(1)) = \frac{x_3 - x_1}{2}\sigma(-1, \alpha, 1) \; ,$$

and similarly,

$$\sigma(x_1, y) + \sigma(x_2, y) + \sigma(x_3, y) = \frac{x_3 - x_1}{2}\big(\sigma(-1, y') + \sigma(\alpha, y') + \sigma(1, y')\big) \; ,$$

where $y' = f^{-1}(y)$. Therefore, inequality (6) reduces to the following:

$$\sigma(-1, \alpha, 1) \leqslant \sigma(-1, y') + \sigma(\alpha, y') + \sigma(1, y'), \quad \forall y' \in \mathbb{R} \, .$$

But

$$\sigma(-1, \alpha, 1) = \frac{1}{3}\sqrt{6 + 2\alpha^2} \leqslant \frac{1}{3}\sqrt{8} < 1 \, ,$$

and on the other side,

$$\sigma(-1, y') + \sigma(\alpha, y') + \sigma(1, y') = \frac{1}{2}\big(|1 + y'| + |\alpha - y'| + |1 - y'|\big) \geqslant 1 \, .$$

□

*Remark 2* • The standard deviation is a contractive multidistance [1]; this means that for any non constant list there exists at least one point which strictly decreases the multidistance when added to the list. The mean of the list can be this point:

$$\sigma(x_1, \ldots, x_n, \bar{x}) = \sqrt{\frac{n}{n+1}}\sigma(x_1, \ldots, x_n) < \sigma(x_1, \ldots, x_n) \, . \tag{7}$$

- The variance fulfills conditions md1, md2. Also, it follows from the previous proof that it fulfills condition md3 for $n \geqslant 3$. But it is not a multidistance because it does not work for $n = 2$. For example, if we take the values 0, 2 and their arithmetic mean 1, we have:

$$\sigma^2(0, 2) = 1 \nleqslant \sigma^2(0, 1) + \sigma^2(1, 2) = 0.25 + 0.25 = 0.5 \, . \tag{8}$$

A class of multidistances, remarkable in this work, is the class of the so called functionally expressible multidistances.

**Definition 4** [4] Let $D$ be a multidistance on a set $X$ and $d$ an ordinary distance on the same set. We will say that $D$ is functionally expressible from $d$ and $F$, or $(d, F)$-functionally expressible, if there exist a symmetric function $F\colon \bigcup_{n\geqslant 1}(\mathbb{R}^+)^n \to \mathbb{R}^+$ such that:

$$D(x_1, \ldots, x_n) = F(d(x_1, x_2), \ldots, d(x_i, x_j), \ldots, d(x_{n-1}, x_n)) \, , \tag{9}$$

for all $n \geqslant 2$, $1 \leqslant i < j \leqslant n$ and $x_1, \ldots, x_n \in X$.

*Example 2* The standard deviation is a multidistance, as shown in Proposition 1, and it can be proved that it is functionally expressible. We do it with the help of the following well-known formula, which expresses the standard deviation in terms of the absolute differences $|x_i - x_j|$, that is, the pairwise distances:

$$\sigma(x_1, \ldots, x_n) = \frac{1}{n}\sqrt{\sum_{i<j} |x_i - x_j|^2} \ . \tag{10}$$

Therefore,

$$\sigma(x_1, \ldots, x_n) = F(|x_1 - x_2|, |x_1 - x_3|, \ldots, |x_{n-1} - x_n|) \ , \tag{11}$$

where the function $F$ should be given by

$$F(a_1, \ldots, a_m) = \sqrt{\frac{2}{1 + 4m + \sqrt{1 + 8m}} \sum_{i=1}^{m} a_i^2} \ . \tag{12}$$

The following example proves the existence of non functionally expressible multidistances.

*Example 3* Consider the function $D\colon \bigcup_{n\geqslant 1}(\mathbb{R}^2)^n \to \mathbb{R}^+$ defined in this way: $D(P_1, \ldots, P_n)$ is the length of the diagonal of the smallest rectangle, with sides parallel to the axes, containing the points $P_1, \ldots, P_n$. Note that the restriction of $D$ to $(\mathbb{R}^2)^2$ is the Euclidean distance $d$.

It can be proved that $D$ is a multidistance. But it is not $d$–functionally expressible: if we take, for example, the points $P_1 = (0, 0)$, $P_2 = (0, 1)$ and $P_3 = (1, 0)$, their pairwise distances are $d(P_1, P_2) = d(P_1, P_3) = 1$, $d(P_2, P_3) = \sqrt{2}$, and their multidistance is $D(P_1, P_2, P_3) = \sqrt{2}$.

But if we change the last two ones to $P_2' = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ and $P_3' = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$, the pairwise distances are the same as before but the multidistance changes: $D(P_1, P_2', P_3') = \sqrt{\frac{5}{2}}$.

Therefore, the value taken by the multidistance is not determined by the pairwise distances and hence, $D$ is not $d$–functionally expressible.

Comparing Definitions 2 and 3 (with $X = \mathbb{R}^k$ and $d$ being the Euclidean distance), it can be observed that conditions $\Delta 1$, $\Delta 2$ are the same as md1, md2. Moreover, condition $\Delta 3$ holds for functionally expressible multidistances, and an additional condition can be given in order to fulfill $\Delta 4$.

**Proposition 2** *Let $D$ be a $(d, F)$-functionally expressible multidistance on $\mathbb{R}^k$. If there exists $c > 0$ such that, for all $n \geqslant 1$, $t_1, \ldots, t_n$ and $a \geqslant 0$, it holds:*

$$F(at_1, \ldots, at_m) = a^c F(t_1, \ldots, t_m) \ , \tag{13}$$

*then $D$ is a dispersion measure.*

*Proof* For any list $(x_1, \ldots, x_n) \in (\mathbb{R}^k)^n$ and any isometry $\phi$ defined on $\mathbb{R}^k$, it holds:

$$\begin{aligned} D(x_1, \dots, x_n) &= F(\dots, d(x_i, x_j), \dots) \\ &= F(\dots, d(\phi(x_i), \phi(x_j)), \dots) \\ &= D(\phi(x_1), \dots, \phi(x_n)) \ . \end{aligned}$$

Hence, $D$ is invariant under isometries.

Also, condition $\Delta 4$ holds, because of (13). □

## 3.1 Sum-Based Multidistances

The multidistances $D_\Sigma^\lambda : \bigcup_{n \geqslant 1} (\mathbb{R}^k)^n \to \mathbb{R}^+$, defined for all $(x_1, \dots, x_n) \in (\mathbb{R}^k)^n$ by:

$$D_\Sigma^\lambda (x_1, \dots, x_n) = \begin{cases} 0 & \text{si } n = 1, \\ \lambda(n) \sum_{i<j} d(x_i, x_j) & \text{si } n \geqslant 2, \end{cases} \tag{14}$$

with $\lambda(2) = 1$ and $0 < \lambda(n) \leqslant \frac{1}{n-1}$ for all $n \geqslant 3$, are said to be sum-based multidistances.

**Proposition 3** *Sum-based multidistances are 1-dispersion measures.*

*Proof* Only conditions $\Delta 3$, $\Delta 4$ must be checked. The first follows from the fact that this kind of multidistances are functionally expressible. The second one is inmediate, taking into account that $d(ax_i, ax_j) = ad(x_i, x_j)$, for all $a \in \mathbb{R}$, and $x_i, x_j$. □

## 3.2 Fermat λ-Multidistances

These multidistances on the set $\bigcup_{n \geqslant 1} (\mathbb{R}^k)^n$ are defined as follows:

$$D_F^\lambda (x_1, \dots, x_n) = \lambda(n) D_F (x_1, \dots, x_n), \tag{15}$$

where $\lambda(2) = 1$ and $\lambda(n) \in (0, 1]$ for all $n \geqslant 3$.

**Proposition 4** *Fermat λ-multidistances are 1-dispersion measures.*

*Proof* If $r$ is the minimum of the sum $\sum_{i=1}^n d(x_i, x)$ $x \in \mathbb{R}^k$. then the minimum of $\sum_{i=1}^n d(\phi(x_i), x)$ and $\sum_{i=1}^n d(ax_i, x)$ are reached at $\phi(r)$ and $ar$, respectively, and so $\Delta 3$, $\Delta 4$ hold. □

### *3.3 OWA-Based Multidistances*

Let $W = (W_n)_{n\geqslant 1}$ be a family of OWA operators. The weights of the OWA $W_n$ will be denoted by $\omega_1^n, \ldots, \omega_n^n$.

An OWA-based multidistance is a function $D_W : \bigcup_{n\geqslant 1}(\mathbb{R}^k)^n \to \mathbb{R}^+$ defined, for all $(x_1, \ldots, x_n) \in (\mathbb{R}^k)^n$, in this way:

$$D_W(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if } n = 1, \\ W_n(\overbrace{d(x_1, x_2), \ldots, d(x_{n-1}, x_n)}^{\binom{n}{2}}) & \text{if } n \geqslant 2, \end{cases} \tag{16}$$

with the weights of the OWA operators of the family $W$ fulfilling this condition:

$$\omega_1^{\binom{n}{2}} + \cdots + \omega_{n-1}^{\binom{n}{2}} > 0, \quad \forall n \geqslant 2\,. \tag{17}$$

We can establish the following result.

**Proposition 5** *OWA-based multidistances on $\mathbb{R}^k$ are 1-dispersion measures.*

*Proof* OWA-based multidistance are obviously functionally expressible, from the expression (17), and so they are invariant under isometries. Also condition $\Delta 4$ holds, due to the fact that OWA operators are homogeneous of degree 1. □

It has been found out that multidistances belonging to these three families are dispersion measures. The versatility of these families, and of multidistances in general, allows choosing the appropriate ones to be used as measures of dispersion in contexts where their character, mainly given by the generalized triangle inequality md3, could be required.

**Acknowledgments** This work has been supported by the Spanish Government projects TIN2013-42795-P and TIN2014-56381-REDT (LODISCO).

## References

1. Calvo T, Martín J, Mayor G (2012) Measures of disagreement and aggregation of preferences based on multidistances. In: Greco S et al (eds) IPMU 2012, Part IV, 549558. Springer, Berlin
2. Kołacz A, Grzegorzewski P (2016) Measures of dispersion for multidimensional data. Eur J Oper Res doi:10.1016/j.ejor.2016.01.011
3. Martín J, Mayor G (2011) Multi-argument distances. Fuzzy Set Syst 167:92–100
4. Martín J, Mayor G, Valero O (2011) Functionally expressible multidistances. Proc EUSFLAT-LFA 2011:41–46
5. Martínez-Panero M, García-Lapresta JL, Meneses LC Multidistances and dispersion measures (to appear)

# Full Conglomerability, Continuity and Marginal Extension

**Enrique Miranda and Marco Zaffalon**

**Abstract** We investigate fully conglomerable coherent lower previsions in the sense of Walley, and some particular cases of interest: envelopes of fully conglomerable linear previsions, envelopes of countably additive linear previsions and fully disintegrable linear previsions. We study the connections with continuity and countable super-additivity, and show that full conglomerability can be characterised in terms of a supremum of marginal extension models.

**Keywords** Coherent lower previsions · Conglomerability · Disintegrability · Marginal extension

## 1 Introduction

Conglomerability of a probability $P$ was first discussed by Bruno de Finetti in [4]. If we consider a partition $\mathcal{B}$ of the possibility space $\Omega$ such that $P(B) > 0$ for every $B \in \mathcal{B}$, conglomerability means that

$$(\forall A \subseteq \Omega) \inf_{B \in \mathcal{B}} P(A|B) \leq P(A) \leq \sup_{B \in \mathcal{B}} P(A|B). \tag{1}$$

A related (but stronger) notion was later studied by Dubins, with the name *disintegrability* [3]. Other studies in the precise case were made in [1, 2, 9, 10].

Imposing as well as checking conglomerability can be technically difficult. Partly for this reason, there are different schools of thought about the previous question: those who reject that conglomerability should be a rationality requirement—among them looms the figure of de Finetti himself; and those who think it should be imposed,

E. Miranda (✉)
University of Oviedo, C-Calvo Sotelo s/n, 33007 Oviedo, Spain
e-mail: mirandaenrique@uniovi.es

M. Zaffalon
IDSIA, Lugano, Switzerland
e-mail: zaffalon@idsia.ch

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_44

often in the light of the paradoxical situations that the lack of conglomerability may lead to. Among the latter stands Peter Walley, who has proposed a behavioural theory of *imprecise* probabilities, where the core modelling unit is a closed convex set of finitely additive probabilities [11]. This theory is essentially Peter Williams' earlier theory of imprecise probability [12] with an additional axiom of conglomerability for sets of probabilities, which coincides with Eq. (1) in the special case of precise probability (and with disintegrability if we require that the conditional model is also precise). The notion of conglomerability is nonetheless not univocally defined in the literature; for this reason, in Sect. 3 we try to sort out the situation by examining and comparing the different proposals in some detail.

In previous papers we have provided a behavioural support for conglomerability [13] and we have showed that it may be a difficult condition to work with in practice [7, 8]. Here we investigate whether at least the notion of *full* conglomerability (that is, conglomerability with respect to every partition) admits a simple treatment. To this end, we make a thorough mathematical study of the properties of full conglomerability and its relations to other notions: continuity (in various forms), countable super-additivity, and marginal extension. Due to limitations of space, the proofs of the results as well as some relevant counterexamples have been omitted.

## 2 Preliminary Notions

Let us introduce the basic elements of the theory of coherent lower previsions. We refer to [11] for more details. Consider a possibility space $\Omega$. A *gamble* is a bounded map $f : \Omega \to \mathbb{R}$. One instance of gambles are the *indicator* gambles of sets $B \subseteq \Omega$, which we shall denote by $I_B$ or $B$. We denote by $\mathcal{L}(\Omega)$ the space of all gambles on $\Omega$.

A *linear prevision* on $\mathcal{L}(\Omega)$ is a linear operator satisfying $P(f) \geq \inf f$ for all $f \in \mathcal{L}(\Omega)$. It is the expectation operator with respect to a finitely additive probability. When its restriction to events is countably additive, meaning that $P(\cup_n B_n) = \sum_n P(B_n)$ for any countable family $(B_n)_n$ of pairwise disjoint events, we say that $P$ is a *countably additive linear prevision*.

A *coherent lower prevision* $\underline{P}$ on $\mathcal{L}(\Omega)$ is the lower envelope of a closed and convex set of linear previsions. The conjugate upper envelope $\overline{P}$ is called a *coherent upper prevision*, and it holds that $\overline{P}(f) = -\underline{P}(-f)$ for all $f$. We let $\mathcal{M}(\underline{P}) := \{P \text{ linear prevision} : (\forall f)\ P(f) \geq \underline{P}(f)\}$ and call it the *credal set* associated with $\underline{P}$. More generally, we say that a map $\underline{P} : \mathcal{L}(\Omega) \to \mathbb{R}$ *avoids sure loss* when it is dominated by some coherent lower prevision. The smallest such prevision is called its *natural extension*, and it coincides with the lower envelope of the non-empty set $\mathcal{M}(\underline{P})$.

A coherent lower prevision is in a one-to-one correspondence with its associated set of *strictly desirable gambles* $\underline{\mathcal{R}} := \{f : \underline{P}(f) > 0 \text{ or } f \gneq 0\}$, in the sense that $\underline{P}(f) = \sup\{\mu : f - \mu \in \underline{\mathcal{R}}\}$ for all $f \in \mathcal{L}(\Omega)$; the closure $\overline{\mathcal{R}}$ of the set of strictly desirable gambles in the topology of uniform convergence is called the set of *almost-desirable gambles*, and it satisfies $\overline{\mathcal{R}} = \{f : \underline{P}(f) \geq 0\}$.

The notion of coherence can also be extended to the conditional case. Let $\mathcal{B}$ be a partition of $\Omega$. A *separately coherent* conditional lower prevision is a map $\underline{P}(\cdot|\mathcal{B}) := \sum_{B\in\mathcal{B}} I_B\underline{P}(\cdot|B)$, and where for every $B \in \mathcal{B}$ the functional $\underline{P}(\cdot|B) : \mathcal{L}(\Omega) \to \mathbb{R}$ is a coherent lower prevision satisfying $\underline{P}(B|B) = 1$.

Given a coherent lower prevision $\underline{P}$ and a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$, they are (jointly) *coherent* when $\underline{P}(G(f|B)) = 0$ for all $f \in \mathcal{L}(\Omega)$, $B \in \mathcal{B}$ and $\underline{P}(G(f|\mathcal{B})) \geq 0$ for all $f \in \mathcal{L}(\Omega)$, where $G(f|B) := B(f - \underline{P}(f|B))$ and $G(f|\mathcal{B}) := \sum_B G(f|B) = f - \underline{P}(f|\mathcal{B})$.

This notion is based on what Walley called the *conglomerative principle*, which means that if a gamble $f$ satisfies that $I_B f$ is desirable for any $B \in \mathcal{B}$, then $f$ should also be desirable. This is the main point of controversy between Walley's and de Finetti's approaches. The latter only requires that a finite sum of desirable gambles is again desirable, and this yields a different notion of conditional coherence, usually referred to as *Williams coherence* [12].

The notion of natural extension can also be considered in the conditional case. Given a coherent lower prevision $\underline{P}$ and a partition $\mathcal{B}$ of $\Omega$, its *conditional natural extension* $\underline{P}(\cdot|\mathcal{B})$ is given by

$$\underline{P}(f|B) := \begin{cases} \inf_B f & \text{if } \underline{P}(B) = 0, \\ \sup\{\mu : \underline{P}(B(f-\mu)) \geq 0\} & \text{otherwise} \end{cases} \tag{2}$$

for any $f \in \mathcal{L}(\Omega)$. It always holds that $\underline{P}(G(f|B)) = 0$ for all $f \in \mathcal{L}(\Omega)$, $B \in \mathcal{B}$, so $\underline{P}$, $\underline{P}(\cdot|\mathcal{B})$ are coherent if and only if $\underline{P}(G(f|\mathcal{B})) \geq 0$ for all $f \in \mathcal{L}(\Omega)$.

## 3 Different Notions of Conglomerability in the Literature

As we mentioned in the Introduction, conglomerability was first introduced by de Finetti in [4] in terms of Eq. (1). The conditional probability $P(A|B)$ in that equation is derived from the unconditional one by Bayes' rule, so that $P(A|B) = P(A \cap B)/P(B)$, whenever $P(B) \neq 0$. However, de Finetti argued [5, Chap. 5] that it also makes sense to consider the conditional probability $P(A|B)$ when the event $B$ has probability 0 but is not deemed impossible. In that case, he suggested to define a *full conditional measure* as that considered in [3, Sect. 3].

There exists a connection between full conditional measures and the theory of coherent previsions: if we represent a full conditional measure on $\mathcal{P}(\Omega) \times (\mathcal{P}(\Omega) \setminus \emptyset)$ as a family of conditional and unconditional assessments $\{P(\cdot|B) : B \subseteq \Omega\}$, then these conditional previsions satisfy the notion of Williams coherence [12, Proposition 6]. On the other hand, as Schervisch, Seidenfeld and Kadane have established in [9, 10], if the linear prevision that results from restricting a full conditional measure to $\mathcal{P}(\Omega)$ is not countably additive, then there is some partition $\mathcal{B}$ of $\Omega$ where Eq. (1) is violated. In other words, under this approach the only fully conglomerable models are the countably additive ones.

On the other hand, Walley [11, Sect. 6.8.1] calls a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\Omega)$ $\mathcal{B}$-*conglomerable* if for any gamble $f$ such that $\underline{P}(Bf) \geq 0$ for all $B \in \mathcal{B}$ with $\underline{P}(B) > 0$, it holds that $\underline{P}(\sum_{\underline{P}(B)>0} Bf) \geq 0$. This is equivalent to the existence of a conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ such that $\underline{P}$, $\underline{P}(\cdot|\mathcal{B})$ are jointly coherent, and also to the coherence of $\underline{P}$ with its conditional natural extension. Thus, conglomerability means that the coherent lower prevision $\underline{P}$ can be updated in a coherent way to a conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$. The notion can be applied in particular to linear previsions. However, in that case we may also require that the linear prevision can be updated into a *linear* model; this gives rise to a stronger notion, called $\mathcal{B}$-*disintegrability*. From [11, Theorem 6.5.7], the $\mathcal{B}$-disintegrability of a linear prevision is equivalent to the existence of a conditional linear prevision $P(\cdot|\mathcal{B})$ such that $P = P(P(\cdot|\mathcal{B}))$.

We say that $\underline{P}$ is *fully conglomerable* when it is $\mathcal{B}$-conglomerable for every partition $\mathcal{B}$ of $\Omega$. In a similar manner, we say that a linear prevision $P$ is *fully disintegrable* when for every partition $\mathcal{B}$ there is some conditional linear prevision $P(\cdot|\mathcal{B})$ such that $P = P(P(\cdot|\mathcal{B}))$.

If a lower prevision $\underline{P}$ is fully conglomerable, then we can define a family of conditional lower previsions $\mathcal{H} := \{\underline{P}(\cdot|\mathcal{B}) : \mathcal{B} \text{ partition of } \Omega\}$ with the property that $\underline{P}$, $\underline{P}(\cdot|\mathcal{B})$ are coherent for every partition $\mathcal{B}$. It can be checked that these conditional lower previsions are also coherent with each other, in the sense that they can all be induced by a common fully conglomerable set of desirable gambles. This means that when we consider the family of all partitions, coherence becomes equivalent to the notion of *conglomerable coherence* studied in much detail in [7]. In the same manner that the natural extension of a lower prevision is the smallest dominating coherent lower prevision, we shall call the *fully conglomerable natural extension* the smallest fully conglomerable coherent lower prevision that dominates $\underline{P}$, in case it exists.

We see then that the two approaches are different, basically because of the manner the problem of conditioning on sets of (lower) probability zero is dealt with. In de Finetti's case, it is advocated to use full conditional measures, while in Walley's case these sets are not taken into account (in the lower prevision approach we are considering here; a more informative model would be that of sets of desirable gambles). In this sense, Walley's condition is close to what Armstrong called *positive conglomerability* in [1]. The different approach means for instance that a linear prevision whose restriction to events is $\{0, 1\}$-valued will always be fully conglomerable for Walley, while it may not be so for de Finetti. Another key difference is in the rejection by de Finetti of the conglomerative principle, that makes the conditional models subject to a different consistency condition (Williams coherence for de Finetti, and the stronger version of Walley in his case).

## 4 Full Conglomerability in the Precise Case

In the precise case, we consider three properties for a linear prevision $P$:

M1. $P$ is countably additive.
M2. $P$ is fully disintegrable.
M3. $P$ is fully conglomerable.

By [11, Theorem 6.9.1], condition M1 implies M2; on the other hand, it follows from its definition that a fully disintegrable linear prevision is in particular fully conglomerable. With respect to the converse implication, we shall consider two cases: linear previsions whose restrictions to events have a finite range (called *molecular* in [2]) and those whose restrictions to events have infinite range (called *non-molecular* in [2]).

**Proposition 1** *Let $P$ be a linear prevision on $\mathcal{L}(\Omega)$.*

1. *If $P$ is molecular, then for every partition $\mathcal{B}$ of $\Omega$, $|\{B \in \mathcal{B} : P(B) > 0\}| < +\infty$, and as a consequence, $P$ is fully conglomerable.*
2. *If $P$ is non-molecular, then it is countably additive if and only if it is fully conglomerable. In that case, $P(\{\omega \in \Omega : P(\omega) > 0\}) = 1$.*

In [9, Theorem 3.3] it is proven that any full conditional measure whose associated unconditional probability is molecular and not countably additive is not fully disintegrable. In other words, countable additivity and full disintegrability are equivalent in the molecular case provided we enter the framework of full conditional measures.

Next we study the connection with continuity. We consider the following continuity conditions:

C1. $(f_n)_{n\in\mathbb{N}} \to f \Rightarrow (\underline{P}(f_n))_{n\in\mathbb{N}} \to \underline{P}(f)$.
C2. $(f_n)_{n\in\mathbb{N}} \downarrow f \Rightarrow (\underline{P}(f_n))_{n\in\mathbb{N}} \downarrow \underline{P}(f)$.
C3. $(f_n)_{n\in\mathbb{N}} \downarrow 0 \Rightarrow (\underline{P}(f_n))_{n\in\mathbb{N}} \downarrow 0$.
C4. $(f_n)_{n\in\mathbb{N}} \uparrow f \Rightarrow (\underline{P}(f_n))_{n\in\mathbb{N}} \uparrow \underline{P}(f)$.

It is not difficult to show the following:

**Proposition 2** *For any linear prevision $P$, M1 $\Leftrightarrow$ C2 $\Leftrightarrow$ C3 $\Leftrightarrow$ C4.*

We deduce from this that condition C1 is sufficient for $P$ to be countably additive. However, it is not necessary. On the other hand, any of the conditions C2–C4 is sufficient for $P$ to be fully disintegrable, and as a consequence also fully conglomerable.

The only open problem at this stage would be the equivalence between M2 and M1. A counterexample would require the definition of a family of conditional linear previsions $\{P(\cdot|\mathcal{B}) : \mathcal{B}$ partition of $\Omega\}$ and an unconditional linear prevision $P$ such that $P = P(P(\cdot|\mathcal{B}))$ for every $\mathcal{B}$ (so $P$ is fully disintegrable) while there exists a finite sub-family of $\{P(\cdot|\mathcal{B}) : \mathcal{B}$ partition of $\Omega\}$ which violates Williams coherence (so that we cannot make a representation in terms of full conditional measures, because if we could, then $P$ would be countably additive by [9]). Such an example seems unlikely, in our opinion.

## 5 Full Conglomerability in the Imprecise Case

In the imprecise case, we consider three properties of a coherent lower prevision $\underline{P}$:

M4. $\underline{P}$ is the lower envelope of a family of countably additive linear previsions.
M5. $\underline{P}$ is the lower envelope of a family of fully conglomerable linear previsions.
M6. $\underline{P}$ is fully conglomerable.

Analogous conditions to M4, M5 (in terms of upper envelopes) can be established for a coherent upper prevision $\overline{P}$. It is immediate to see that

$$\text{M1} \Rightarrow \text{M3} \Rightarrow \text{M5} \Rightarrow \text{ M6 and M1 } \Rightarrow \text{M4} \Rightarrow \text{M5} \Rightarrow \text{M6}.$$

However, the remaining implications do not hold: on the one hand, a linear prevision may be fully conglomerable without being countably additive; moreover, there are fully conglomerable coherent lower previsions that are not dominated by any fully conglomerable (and as consequence by any countably additive) linear prevision [11, Example 6.9.6].

With respect to M4, Krätschmer established in [6, Sect. 5] that a 2-alternating upper probability on $\mathcal{P}(\Omega)$ is the upper envelope of a family of countably additive probabilities if and only if $\overline{P}(A) = \sup\{\overline{P}(B) : A \supseteq B \text{ finite}\}$ for every $A \subseteq \Omega$. However, we have shown that the above condition does not characterise M4 in general. Nevertheless, we can give a necessary and sufficient condition in the particular case where $\Omega = \mathbb{N}$:

**Proposition 3** *Let $\overline{P}$ be a coherent upper prevision on $\mathcal{L}(\mathbb{N})$. Then $\overline{P}$ satisfies M4 $\Leftrightarrow (\forall n \in \mathbb{N})\ \overline{P} = \sup \mathcal{M}_n \Leftrightarrow (\forall f \geq 0)\ \overline{P}(f) = \lim_n \overline{P}(f I_{\{1,\ldots,n\}}) \Leftrightarrow (\forall f \geq 0)\quad \overline{P}(f) = \sup\{\overline{P}(g) : g \leq f, \mathit{supp}(g) \text{ finite}\}$, where $\mathcal{M}_n := \{P \leq \overline{P} : \lim_m P(\{1,\ldots,m\}) \geq 1 - \frac{1}{n}\}$ and $(\forall g)\ \mathit{supp}(g) = \{n : g(n) \neq 0\}$.*

Next, we study the connection with the continuity properties C1–C4. On the one hand, we deduce from the precise case that none of them is necessary for $\underline{P}$ to belong to M5, M6. On the other hand, we have that:

**Proposition 4** *C1 $\Rightarrow$ C4 $\Rightarrow$ M4 $\Rightarrow$ C2, M5 $\Rightarrow$ M6 and C2 $\Rightarrow$ C3. Moreover, no additional implication other than the ones that immediately follow from these holds.*

Next we investigate the connection with the following condition:

M7. $(\forall (f_n)_n \subseteq \mathcal{L}(\Omega)$ such that $\sum_n f_n \in \mathcal{L}(\Omega))\ \underline{P}\left(\sum_n f_n\right) \geq \sum_n \underline{P}(f_n)$.

The reason for our investigation is that both countable super-additivity and conglomerability are quite related to the closedness of a set of desirable gambles under countable sums. Specifically, we have proven the following:

**Proposition 5** *Let $\underline{P}$ be a coherent lower prevision and let $\underline{\mathcal{R}}$, $\overline{\mathcal{R}}$ denote its associated sets of strictly desirable and almost desirable gambles, respectively. Then each of the following statements implies the next:*

1. *$\underline{P}$ satisfies M7.*
2. $(\forall (f_n)_n \subseteq \underline{\mathcal{R}} : \sum_n f_n \in \mathcal{L}(\Omega))\ \sum_n f_n \in \underline{\mathcal{R}}$.
3. $(\forall (f_n)_n \subseteq \underline{\mathcal{R}} : \sum_n f_n \in \mathcal{L}(\Omega))\ \sum_n f_n \in \overline{\mathcal{R}}$.
4. *$\underline{P}$ satisfies C3.*

The connection between M7 and the other conditions is given by

$$\text{C2} \Rightarrow \text{M7} \Rightarrow\ \text{C3 and M7}\ \Rightarrow \text{M6},$$

together with those derived from Proposition 4. We deduce that if $P$ is linear,

$$\text{C1} \Rightarrow \text{M1} \Leftrightarrow \text{C2} \Leftrightarrow \text{M7} \Leftrightarrow \text{C3} \Leftrightarrow \text{C4} \Rightarrow \text{M2} \Rightarrow \text{M3}.$$

The only open problem left at this stage is whether M7 and C2 are equivalent.

## 6 Full Conglomerability and Marginal Extension

From [11, Theorem 6.8.2], given a coherent lower prevision $\underline{P}$ and a partition $\mathcal{B}$ of $\Omega$, it holds that $\underline{P}$ is $\mathcal{B}$-conglomerable if and only if $\underline{P} \geq \underline{P}(\underline{P}(\cdot|\mathcal{B}))$, where $\underline{P}(\cdot|\mathcal{B})$ is the conditional natural extension of $\underline{P}$, given by Eq. (2). Thus, $\underline{P}$ is fully conglomerable if and only if $\underline{P} \geq \sup_{\mathcal{B}\text{ partition}} \underline{P}(\underline{P}(\cdot|\mathcal{B})) := \underline{Q}$.

The concatenation $\underline{P}(\underline{P}(\cdot|\mathcal{B}))$ of a marginal and a conditional lower prevision is called a *marginal extension model* [11, Sect. 6.7]; this is an extension of the product rule to the imprecise case. The condition above tells us then that fully conglomerable lower previsions are always the supremum of a family of marginal extension models. Our next proposition summarizes the relationship between $\underline{P}$ and the functional $\underline{Q}$ it determines:

**Proposition 6** *Let $\underline{P}$ be a coherent lower prevision and $\underline{F}$ its fully conglomerable natural extension (if it exists), and define $\underline{Q}$ as above.*

1. $\underline{P} \leq \underline{Q} \leq \underline{F}$.
2. *$\underline{P}$ is fully conglomerable $\Leftrightarrow$ $\underline{P} = \underline{Q}$.*
3. *$\underline{Q}$ does not avoid sure loss in general, and $\mathcal{M}(\underline{Q}) \neq \emptyset \not\Rightarrow \underline{P}$ satisfies M6.*

Thus, the full conglomerability of $\underline{P}$ implies the coherence of $\underline{Q}$. Although it is an open problem whether the converse holds in general, it is easy to see that when $P$ is linear, then $\underline{Q} \geq P$ is coherent if and only if $\underline{Q} = P$ (it cannot be that $\underline{Q}(f) > P(f)$ and still be that $\underline{Q}$ is coherent), so in the precise case we have the equivalence.

## 7 Conclusions

Our results show that countably additive models and their envelopes seem to be easier to use in practice than fully conglomerable ones; although the connection with continuity in the precise case is well known, as it follows almost immediately from existing results from probability theory, in the imprecise case we have given a necessary and a sufficient condition, as well as a characterisation in terms of the natural extension from gambles with a finite range. In our view, this indicates that envelopes of countably additive linear previsions may be more interesting in practice, and they could be a tool to guarantee the property of full conglomerability.

The definition of joint coherence of a conditional and an unconditional lower prevision has led us to define the functional $\underline{Q}$ as a supremum of marginal extensions. A deeper study of this functional is one of the main open problems for future work; in particular, we would like to determine whether the existence of the fully conglomerable natural extension is equivalent (and not only sufficient) to $\underline{Q}$ avoiding sure loss, and whether the coherence of $\underline{Q}$ is sufficient (and not only necessary) for its equality with the fully conglomerable natural extension of $\underline{P}$.

More generally, it would be interesting to make a deeper comparison between our results and the ones established by Seidenfeld et al. for the precise case by means of full conditional measures.

**Acknowledgments** The research reported in this paper has been supported by project TIN2014-59543-P.

## References

1. Armstrong T (1990) Conglomerability of probability measures on Boolean algebras. J Math Anal Appl 150:335–358
2. Armstrong T, Prikry K (1982) The semi-metric on a Boolean algebra induced by a finitely additive probability measure. Pac J Math 99:249–264
3. Dubins L (1975) Finitely additive conditional probabilities, conglomerability and disintegrations. Ann Prob 3:88–99
4. de Finetti B (1930) Sulla proprietà conglomerativa delle probabilità subordinate. Rend Real Inst Lomb 63:414–418
5. de Finetti B (1972) Probability, induction and statistics. Wiley, London
6. Krätschmer V (2003) When fuzzy measures are upper envelopes of probability measures. Fuzzy Sets Syst 138:455–468
7. Miranda E, Zaffalon M (2013) Conglomerable coherence. Int J Appl Reas 54:1322–1350
8. Miranda E, Zaffalon M, de Cooman G (2012) Conglomerable natural extension. Int J Appl Reas 53:1200–1227
9. Schervisch M, Seidenfeld T, Kadane J (1984) The extent of nonconglomerability of finitely additive probabilities. Zeit Wahr Verw Geb 66:205–226
10. Seidenfeld T, Schervisch M, Kadane J (1998) Non-conglomerability for finite-valued finitely additive probability. Sank 60:476–491
11. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, London
12. Williams P (2007) Notes on conditional previsions. Int J Appl Reas 44:366–383
13. Zaffalon M, Miranda E (2013) Artif Int 198:1–51

# On Extreme Points of p-Boxes and Belief Functions

**Ignacio Montes and Sebastien Destercke**

**Abstract** The extreme points of convex probability sets play an important practical role, especially as a tool to obtain specific, easier to manipulate sets. Although this problem has been studied for many models (probability intervals, possibility distributions), it remains to be studied for imprecise cumulative distributions (a.k.a. p-boxes). This is what we do in this paper, where we characterize the maximal number of extreme points of a p-box, give a family of p-boxes that attains this number and show an algorithm that allows to compute the extreme points of a given p-box. To achieve all this, we also provide what we think to be a new characterization of extreme points of a belief function.

## 1 Introduction

Imprecise probability theory [11] is a powerful unifying framework for uncertainty treatment, relying on convex sets of probabilities, or *credal sets*, to model the uncertainty. Formally, they encompass many existing models: belief functions, possibility distributions, probability intervals, .... To apply such models, it is important to study their practical aspects, among which is the characterization of their extreme points. Indeed, these extreme points can be used in many settings, such as graphical models or statistical learning.

I. Montes
Universidad Carlos III de Madrid, Madrid, Spain
e-mail: igmontes@est-econ.uc3m.es

S. Destercke (✉)
CNRS, UMR Heudiasyc, Universite de Technologie de Compiegne,
Compiegne, France
e-mail: sebastien.destercke@hds.utc.fr

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_45

Extreme points of many models have already been studied. For instance, Dempster [3] shows that the maximal number of extreme points of a belief function on a $n$-element space[1] is $n!$, and this upper bound was also given for less restrictive models in [12]. It was later [6] proved that the maximal number of extreme points for possibility distributions in a $n$-element space is $2^{n-1}$, and in [8] an algorithm to extract them was provided. In [2], authors studied the extreme points of probability intervals.

One practical and popular model for which extreme points have not been characterized are p-boxes [4]. They are special kinds of belief functions whose focal elements are ordered intervals [5, 10, 11], and are quite instrumental in applications such as risk and reliability analysis.

In this paper, we investigate extreme points of p-boxes: we demonstrate that their maximal number is the Pell number, and give the family of p-boxes for which this bound is obtained. To do so, we introduce a new way to characterize the extreme points of a belief function. We also provide an algorithm to compute the extreme points of a given p-box. Section 2 introduces the new characterization, while Sect. 3.2 studies the extreme points of p-boxes. Due to space restrictions, proofs and side results have been removed.

## 2 Extreme Points of Belief Functions

Given a space $\mathcal{X} = \{x_1, \ldots, x_n\}$, a *basic probability assignment* (bpa) is a function $m : \mathcal{P}(\mathcal{X}) \to [0, 1]$ satisfying $m(\emptyset) = 0$ and $\sum_{B \subseteq \mathcal{X}} m(B) = 1$. A bpa $m$ defines a *belief* Bel and a *plausibility* Pl function [9] by:

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B) \quad \text{and} \quad \mathrm{Pl}(A) = \sum_{B : A \cap B \neq \emptyset} m(B) \quad \forall A \subseteq \mathcal{X}.$$

These two functions are conjugate since $\mathrm{Bel}(A) = 1 - \mathrm{Pl}(A^c)$, and we can focus on one of them. A *focal set* of the belief function Bel is a set $E$ such that $m(E) > 0$, and $\mathcal{F}$ will denote the set of focal sets. A belief function also induces a credal set

$$\mathcal{M}(\mathrm{Bel}) = \{P \text{ Prob.} \mid \mathrm{Bel}(A) \leq P(A)\ \forall A \subseteq \mathcal{X}\}.$$

Being convex, the set $\mathcal{M}(\mathrm{Bel})$ can be characterized by its extreme points,[2] that we will denote $\mathcal{E}xt(\mathrm{Bel})$. It is known [1, 3] that there is a correspondence between the extreme points of a belief function and the permutations of the elements of $\mathcal{X}$.

[1] For the sake of simplicity, we use the terminology "extreme points of a belief function" to refer to the extreme points of the credal set associated with the belief function.

[2] Recall that an extreme point $P$ of $\mathcal{M}(\mathrm{Bel})$ is a point such that, if $P_1, P_2 \in \mathcal{M}(\mathrm{Bel})$ and $\alpha P_1 + (1 - \alpha)P_2 = P$ for some $\alpha \in (0, 1)$, then $P_1 = P_2 = P$.

The extreme point $P_\sigma \in \mathcal{E}xt(\mathrm{Bel})$ associated with the permutation $\sigma$ of $\{1, \dots, n\}$ is given by

$$P_\sigma(\{x_{\sigma(i)}\}) = \mathrm{Bel}(\{x_{\sigma(i)}, \dots, x_{\sigma(n)}\}) - \mathrm{Bel}(\{x_{\sigma(i+1)}, \dots, x_{\sigma(n)}\}) \tag{1a}$$

$$= \sum_{E \subseteq A_i^\sigma} m(E) - \sum_{E \subseteq A_{i+1}^\sigma} m(E) = \sum_{x_{\sigma(i)} \in E, E \cap A_i^{\sigma,C} = \emptyset} m(E) \tag{1b}$$

where $A_i^\sigma = \{x_{\sigma(i)}, \dots, x_{\sigma(n)}\}$ and $A_i^{\sigma,C} = \{x_{\sigma(1)}, \dots, x_{\sigma(i-1)}\}$ is its complement, and the convention $A_{n+1}^\sigma = A_1^{\sigma,C} = \emptyset$. However, we may have that $P_{\sigma_1} = P_{\sigma_2}$, as in general not all permutation give rise to different extreme points, otherwise every belief function would have $n!$ extreme points. Equation (1b) tells us that an extreme point is built iteratively, according to Algorithm 1.

**Algorithm 1:** Extreme point computation

```
Input: $\sigma$, $(Bel)$, $\mathcal{E} = \mathcal{F}$
Output: $P_\sigma$
for $k=1,\dots,n$ do
    For all $E \in \mathcal{E}$ s.t. $x_{\sigma(k)} \in E$, assign $m(E)$ to $P_\sigma(\{x_{\sigma(k)}\})$;
    $\mathcal{E} \leftarrow \mathcal{E} \setminus \{E \in \mathcal{E} | x_{\sigma(k)} \in E\}$
end
```

Let us now introduce another way to characterize this extreme point. To do so, we will denote by $\overline{v}_{i\setminus A} = |\{E \in \mathcal{F} | x_i \in E, E \cap A = \emptyset\}|$ the number of focal sets having $x_i$ as an element and having an empty intersection with $A$. Given a permutation $\sigma$, we denote by $\mathbf{v}^\sigma = (v_1^\sigma, \dots, v_n^\sigma)$ the vector such that

$$v_i^\sigma = \overline{v}_{i \setminus A^{\sigma,C}_{\sigma^{-1}(i)}} = |\{E \in \mathcal{F} | x_i \in E, E \cap \{x_{\sigma(1)}, \dots, x_{\sigma(\sigma^{-1}(i)-1)}\} = \emptyset\}| \tag{2}$$

and by $\mathcal{V}(\mathrm{Bel})$ the set of vectors obtained for all permutation. We will also denote $\overline{\mathbf{v}}_A = (\overline{v}_{1\setminus A}, \dots, \overline{v}_{n\setminus A})$. We then have the following result.

**Proposition 1** *Given Bel, if two permutations $\sigma_1, \sigma_2$ satisfy $P_{\sigma_1} = P_{\sigma_2}$, then $\mathbf{v}^{\sigma_1} = \mathbf{v}^{\sigma_2}$.*

Also note that any vector $\mathbf{v} \in \mathcal{V}(\mathrm{Bel})$ can be associated with a permutation $\sigma$ generating an extreme point (to see this, note the link between Eqs. (2) and (1b)), for instance the permutation having generated it. Since by contraposition of Proposition 1, $\mathbf{v}^{\sigma_1} \neq \mathbf{v}^{\sigma_2}$ implies $P_{\sigma_1} \neq P_{\sigma_2}$, $\mathcal{V}(\mathrm{Bel})$ is in bijection with $\mathcal{E}xt(\mathrm{Bel})$ (any vector induces one and only one distinct extreme point). Given a vector $\mathbf{v} \in \mathcal{V}(\mathrm{Bel})$, we can easily determine a permutation generating it by using Algorithm 2.

**Algorithm 2:** Permutation generating algorithm

```
Input: $\mathbf{v} \in \mathcal{V}(\text{Bel})$, $\mathcal{E} = \mathcal{F}$
Output: One permutation $\sigma$ generating $\mathbf{v}$
for $k=1,\dots,n$ do
    Define $\mathbf{v}$ s.t. $v_i = |\{E \in \mathcal{E} | x_i \in E\}|$ ;
    Find $i$ s.t. $v_i = \overline{v}_{i \setminus A_k^{\sigma,C}}$ /* Getting $A_k^{\sigma,C}$ only requires $\sigma(k-1)$ */
    Define $\sigma(k) = i$;
    $\mathcal{E} \leftarrow \mathcal{E} \setminus \{E \in \mathcal{E} | x_{\sigma(k)} \in E\}$
end
```

*Example 1* Consider a belief function Bel defined on $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ with focal sets $E_1 = \{x_1, x_2\}$, $E_2 = \{x_2, x_3, x_4\}$, $E_3 = \{x_3\}$ and masses 0.2, 0.5 and 0.3, respectively. Consider for example the permutation $\sigma = (1, 2, 3, 4)$. It generates the extreme point $P_\sigma = (0.2, 0.5, 0.3, 0)$. Indeed, according to Algorithm 1, $m(E_1)$ is assigned to $x_1$, $m(E_2)$ to $x_2$ and $m(E_3)$ to $x_3$. Then, $\sigma$ generates the vector $\mathbf{v} = (1, 1, 1, 0)$. Algorithm 2 can then generate permutations (1, 2, 4, 3) or (1, 2, 3, 4), as in the first iteration we only have $v_1 = \overline{v}_{1 \setminus A_1^{\sigma,C}} = 1$, meaning $\sigma(1) = 1$, and in the second iteration we only have $v_2 = \overline{v}_{2 \setminus A_2^{\sigma,C}} = 1$, and so on ....

The extreme points of the belief function in this example, as well as the permutations that generate them, can be seen in Table 1.

Note that this new characterization in terms of "counting" vectors allows us to derive new results about the extreme points of belief functions.

**Proposition 2** *Let Bel be a belief function on $\mathcal{X} = \{x_1, \dots, x_n\}$. The number of extreme points of Bel is $n!$ if and only if $\{x_i, x_j\}$ is a focal set for any $i, j \in \{1, \dots, n\}$ such that $i \neq j$.*

This proposition tells us when a belief function attains the maximal number of extreme points, $n!$. Somehow surprising, the number of focal sets does not matter to

**Table 1** Extreme points of the belief function of Example 1

| Permutation | Probability | $(v_1^\sigma, v_2^\sigma, v_3^\sigma, v_4^\sigma)$ |
|---|---|---|
| (1, 2, 3, 4) (1, 2, 4, 3) | $P_{\sigma_1} = (0.2, 0.5, 0.3, 0)$ | (1, 1, 1, 0) |
| (1, 3, 2, 4) (1, 3, 4, 2) (3, 4, 1, 2) (3, 1, 2, 4) (3, 1, 4, 2) | $P_{\sigma_2} = (0.2, 0, 0.8, 0)$ | (1, 0, 2, 0) |
| (1, 4, 3, 2) (1, 4, 2, 3) (4, 3, 1, 2) (4, 1, 3, 2) (4, 1, 2, 3) | $P_{\sigma_3} = (0.2, 0, 0.3, 0.5)$ | (1, 0, 1, 1) |
| (2, 3, 1, 4) (2, 3, 4, 1) (2, 4, 1, 3) (2, 1, 3, 4) (2, 1, 4, 3) (2, 4, 3, 1) | $P_{\sigma_4} = (0, 0.7, 0.3, 0)$ | (0, 2, 1, 0) |
| (3, 2, 1, 4) (3, 2, 4, 1) (3, 4, 2, 1) | $P_{\sigma_5} = (0, 0.2, 0.8, 0)$ | (1, 0, 2, 0) |
| (4, 3, 2, 1) (4, 2, 3, 1) (4, 2, 1, 3) | $P_{\sigma_6} = (0, 0.2, 0.3, 0.5)$ | (0, 1, 1, 1) |

attain the maximal number of extreme points, the only relevant focal sets are those of cardinality two.

**Proposition 3** *Let Bel be a belief function on $\mathcal{X} = \{x_1, \dots, x_n\}$. Denote by $\mathcal{F}$ the family of focal sets of Bel. Let Bel$'$ be another belief function and let $\mathcal{F}' = \mathcal{F} \cup \{E\}$ be the family of focal sets of Bel$'$, where $E \notin \mathcal{F}$. Then, Bel$'$ has at least as many extreme points as Bel.*

## 3 Extreme Points of p-Boxes

Before studying the extreme points of p-boxes, we need to make a small, useful digression about a specific number sequence: the Pell numbers. Pell numbers form a sequence that follows a recursive relation $\mathcal{P}_0 = 0, \quad \mathcal{P}_1 = 1, \quad \mathcal{P}_n = \mathcal{P}_{n-2} + 2\mathcal{P}_{n-1}$. The first numbers are: 0, 1, 2, 5, 12, 29, 70, .... It is known that $2^{n-1} \leq \mathcal{P}_n \leq n!$ for any $n \geq 1$. As we shall see, it turns out that the maximal number of extreme points of p-boxes on $\mathcal{X}$ is $\mathcal{P}_n$.

### *3.1 Basic Definitions*

From now on we consider a totally ordered set $\mathcal{X} = \{x_1, \dots, x_n\}$ such that $x_1 < \dots < x_n$. A *probability box* or p-box [4] $(\underline{F}, \overline{F})$ is a pair of cumulative distribution functions $\underline{F}, \overline{F} : \mathcal{X} \to [0, 1]$ such that $\underline{F} \leq \overline{F}$. Here we interpret p-boxes as lower and upper bounds of an ill-known cumulative distribution, that induce a credal set

$$\mathcal{M}(\underline{F}, \overline{F}) = \{P \text{ Prob.} \mid \underline{F}(x) \leq F_P(x) \leq \overline{F}(x)\ \forall x \in \mathcal{X}\},$$

where $F_P$ denotes the cumulative distribution function associated with the probability $P$.

It is known that p-boxes are particular instances of belief functions (see [10, 11] for details). That is, to any p-box we can associate a belief function such that $\mathcal{M}(\text{Bel}) = \mathcal{M}(\underline{F}, \overline{F})$. The focal sets $E_1, \dots, E_k$ of this belief function are known to be intervals[3] ordered with respect to the order $\preceq$ between intervals such that

$$[a_1, a_2] \preceq [b_1, b_2] \Leftrightarrow a_1 \leq b_1, a_2 \leq b_2.$$

That is, $E_1 \prec E_2 \prec \dots \prec E_k$. For the reader interested in the way such focal sets can be built, we refer to [5]. This is also a characterization, as any belief function whose focal sets are ordered intervals will be equivalent to a p-box.

[3] By interval, we mean that all elements between $\min E$ and $\max E$ are included in $E$.

### 3.2 Extreme Points of a p-Box

We can easily provide first bounds over the number of extreme points of p-boxes.

**Proposition 4** *The maximal number of extreme points of a p-box on* $\mathcal{X} = \{x_1, \dots, x_n\}$ $(n > 2)$ *lies in the interval* $[2^{n-1}, n!)$.

The exact maximal number of extreme points of a p-box is reached for the following family of p-boxes: the *Pell* p-boxes on $\mathcal{X} = \{x_1, \dots, x_n\}$ are those whose focal sets are

$$\{x_1\},\ \{x_n\},\ \{x_1, x_2\},\ \{x_{n-1}, x_n\},$$
$$\forall i = 2, \dots, n-1,\ \{x_{i-1}, x_i, x_{i+1}\},\ \text{and either } [x_{i-1}, x_{i+2}] \text{ or } [x_i, x_{i+1}].$$

**Theorem 1** *If* $(\underline{F}, \overline{F})$ *is a p-box of the Pell family on* $\mathcal{X} = \{x_1, \dots, x_n\}$*, its number of extreme points is the Pell number* $\mathcal{P}_n$.

**Theorem 2** *The maximal number of extreme points of a p-box defined on* $\mathcal{X} = \{x_1, \dots, x_n\}$ *is the Pell number* $\mathcal{P}_n$*, and is reached if and only if the p-box is of the Pell family.*

### 3.3 Counting the Number of Extreme Points of a p-Box

In this section, we provide an algorithm to enumerate the extreme points of a given p-box. This algorithm builds up a tree by incrementally assigning values $v_i$ to vectors $\mathbf{v} \in \mathcal{V}(\text{Bel})$ as well as corresponding probability values. The $i$th level of the tree corresponds to $v_i$ values, and each leaf then corresponds to a distinct extreme point (whose values can be found back by going from the leaf to the root). Pseudo-Algorithm 3 describes how children are created from a node having depth $d < n$. At a given depth $d$, a node is created (Loop 4–14 of Algorithm 3) for each possible number of focal elements that affect their masses to $x_{d+1}$ (including 0), and the created node receive the corresponding probability $P(x_{d+1})$, the value $v_{d+1}$ of the corresponding permutation vector in $\mathcal{V}$, and the update set of focal elements determining which mass remains to be distributed to which elements. The whole tree can then be built by applying this method recursively, until a depth $n$ is reached. The root node (level 0) simply starts with $\mathcal{E} = \mathcal{F}$.

**Algorithm 3:** Tree building algorithm

```
   Input: Tree node with depth $d < n$ and associated set $\mathcal{E}$ of focal elements
   Output: Children of node
1  $\underline{Nb} \leftarrow \begin{cases} 0 & \text{if } \{x_{d+1}\} \notin \mathcal{E}, \\ 1 & \text{else.} \end{cases}$ ;
2  $\overline{Nb} \leftarrow |\{E_k \in \mathcal{E} | x_{d+1} \in E_k\}|$ /* Number of focal sets containing $x_{d+1}$ */ ;
3  $\underline{k} \leftarrow \inf_{E_k \in \mathcal{E}} k$ ;
4  for $i = \underline{Nb}, \dots, \overline{Nb}$ do
5      $P(x_{d+1}) \leftarrow \sum_{j=\underline{Nb}}^{i} m(E_{j+\underline{k}-1})$          /* $m(E_0) = 0$ */ ;
6      $v_{d+1} \leftarrow i$ ;
7      $\ell^* \leftarrow \max_\ell \{x_\ell \in E_{i+\underline{k}-1}\}$          /* $\ell^* = d + 1$ if $E_{\underline{k}-1}$ */;
8      $\mathcal{E}^* \leftarrow \mathcal{E}$;
9      foreach $E \in \mathcal{E}^*$ such that $x_{d+1} \in E$ do
10         $E \leftarrow E \setminus \{x_1, \dots, x_{\ell^*}\}$ ;
11         if $E = \emptyset$ then Remove $E$ from $\mathcal{E}^*$
12     end
13     foreach $E \in \mathcal{E}^*$ such that $x_{d+1} \notin E$ do
14         if $E \setminus \{x_1, \dots, x_{\ell^*}\} \neq \emptyset$ then $E \leftarrow E \setminus \{x_1, \dots, x_{\ell^*}\}$
15     end
16     Create children of depth $d + 1$ and associate $P(x_{d+1})$, $v_{d+1}$, $\mathcal{E}^*$ to it. ;
17 end
```

*Example 2* Consider a p-box $(\underline{F}, \overline{F})$ on $\{x_1, x_2, x_3, x_4\}$ whose focal sets are given by:

| | $E_1 = \{x_1\}$ | $E_2 = \{x_1, x_2, x_3\}$ | $E_3 = \{x_1, x_2, x_3, x_4\}$ | $E_4 = \{x_3, x_4\}$ |
|---|---|---|---|---|
| $m$ | 0.2 | 0.1 | 0.4 | 0.3 |

Following Algorithm 3 and starting at the root (level 0), at the first step we have $\underline{Nb} = 1$, $\overline{Nb} = 3$, therefore the first level of the tree has three nodes (the root has three children). For $v_1 = 3$, $P(\{x_1\}) = 0.7$, the update gives $\mathcal{E}^* = E_4 = \{x_3, x_4\} = 0.3$, which is used to generate the node children. At the next level, only one node is generated with $v_2 = 0$, $P(\{x_2\}) = 0$, as $\underline{Nb} = \overline{Nb} = 0$ (no focal set contains $x_2$), with $\mathcal{E}^* = E_4 = \{x_3, x_4\} = 0.3$. This node in turns generates two nodes, as $\underline{Nb} = 0$ and $\overline{Nb} = 1$, and so on.

Figure 1 illustrates the process in a synthetic way (as not all details are given, due to lack of space), as well as the extreme points corresponding to leaves of the trees. The development of the second level of the tree is given only for $v_1 = 1$, to illustrate the update of $\mathcal{E}$ (Lines 9–15 of Algorithm 3).

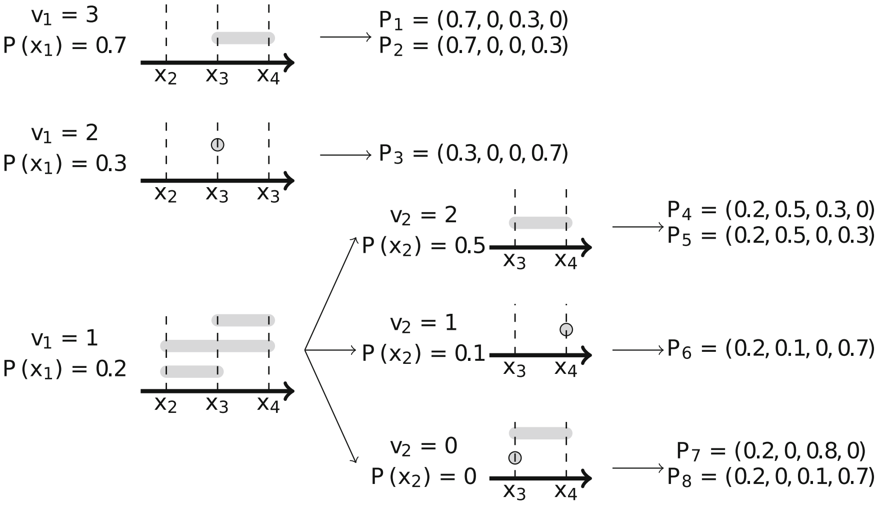


**Fig. 1** Algorithm for extracting the extreme points of Example 2

## 4 Conclusions

In this paper, we have characterized the maximal number of extreme points and have provided an algorithm to enumerate them by means of the construction of a tree structure.

There are still some interesting open problems, for instance we could try to extend our present results to the multivariate case (bivariate p-boxes) [7]. Nevertheless, this seems to be a hard problem because the connection between (univariate) p-boxes and belief functions no longer holds in the bivariate case.

**Acknowledgments** We thank Georg Schollmeyer for insightful exchanges and discussion about this problem. The research reported in this paper has been supported by project TIN2014-59543-P and by LABEX MS2T (ANR-11-IDEX-0004-02).

## References

1. Chateauneuf A, Jaffray JY (1989) Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. Math Soc Sci 17(3):263–283
2. De Campos LM, Huete JF, Moral S (1994) Probability intervals: a tool for uncertain reasoning. Int J Uncertain Fuzziness Knowle-Based Syst 2(2):167–196
3. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–339
4. Ferson S, Kreinovich V, Ginzburg L, Myers DS, Sentz K (2003) Constructing probability boxes and Dempster-Shafer structures. Technical Report SAND2002-4015, Sandia National Laboratories

5. Kriegler E (2005) Utilizing belief functions for the estimation of future climate change. Int J Approx Reason 39(2–3):185–209
6. Miranda E, Couso I, Gil P (2003) Extreme points of credal sets generated by 2-alternating capacities. Int J Approx Reason 33(1):95–115
7. Pelessoni R, Vicig P, Montes I, Miranda E (2015) Bivariate p-boxes. Int J Uncertain Fuzziness Knowl-Based Syst (accepted)
8. Schollmeyer G (2015) On the number and characterization of the extreme points of the core of necessity measures on finite spaces. In: Augustin T, Doria S, Miranda E, Quaeghebeur E (eds) Proceedings of the 9th ISIPTA
9. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton, NJ
10. Troffaes MCM, Destercke S (2011) Probability boxes on totally preordered spaces for multivariate modelling. Int J Approx Reason 52(6):767–791
11. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, London
12. Wallner A (2007) Extreme points of coherent probabilities in finite spaces. Int J Approx Reason 44:339–357

# Modelling the Dependence in Multivariate Longitudinal Data by Pair Copula Decomposition

**Marta Nai Ruscone and Silvia Angela Osmetti**

**Abstract** The aim of the work is to propose a new flexible way of modeling the dependence between the components of non-normal multivariate longitudinal-data by using the copula approach. The presence of longitudinal data is increasing in the scientific areas where several variables are measured over a sample of statistical units at different times, showing two types of dependence: between variables and across time. We propose to model jointly the dependence structure between the responses and the temporal structure of each processes by pair copula contruction (PCC). The use of the copula allows the relaxation of the assumption of multinormality that is typical of the usual model for multivariate longitudinal data. The use of PCC allows us to overcome the problem of the multivariate copulae used in the literature which suffer from rather inflexible structures in high dimension. The result is a new extremly flexible model for multivariate longitudinal data, which overcomes the problem of modeling simultaneous dependence between two or more non-normal responses over time. The explanation of the methodology is accompanied by an example.

## 1 Introduction

Longitudinal data show an increasing occurrence in many scientific research areas where several response variables are measured with reference to a sample of statistical units at different times. The advantage of this study is that it can provide information about subject change, by collecting repeated measurements over time. In this type of data, there are two types of dependence: between variables and over time. The multivariate longitudinal models usually considered in the literature are based on the normality assumption (e.g. [8, 9]). Unfortunately, the empirical evidence shows that normality is certainly not a rule in practice. When the responses are not normal

M. Nai Ruscone (✉)
Università Cattaneo LIUC, Castellanza, VA, Italy
e-mail: mnairuscone@liuc.it

S.A. Osmetti
Università Cattolica del Sacro Cuore, Milano, MI, Italy
e-mail: silvia.osmetti@unicatt.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_46

or when their marginal distributions are not in the same family, alternatives to the multivariate normal models must be found. In order to relax the assumption of normality we introduce the use of the copula function to jointly model the dependence structure between the variables and the temporal structure of each process in the model. In particular, we propose a new model for multivariate non normal longitudinal data based on a D-vine copula that is one of a wider class of vine decompositions recently discussed in the context of graphical models (see [2]). We choose the D-vine copula approach because it is an extremely flexible representation of a multivariate distribution that uses bivariate copula (pair-copula) in a hierarchical manner. Smith et al. [6] use D-vine copula to model the temporal dependence in `univariate longitudinal data` (one variable observed for some subject over time). For multivariate time series ($T$ observations of a $R$-dimensional vector) Smith [7] suggest modeling nonlinear serial and cross-sectional dependence by D-vine copula model. In particular Smith reorder the observations of the multivariate series into the univariate series of dimension $T * R$ and models the joint distribution of the entire series by a D-vine copula of dimension $T * R$. The component pair-copula densities in the D-vine density are grouped together in blocks of pair-copulae used to isolate cross-sectional and serial dependence of the multivariate series. Instead in our work we model a `multivariate longitudinal data` ($T$ observations of a $R$-dimensional vector for a sample of n subject) by using a different D-vine copula approach. First we suppose that the dependence between the responses doesn't depend on the time. The proposed model considers two different levels of analysis. Firstly each longitudinal series, corresponding to a given response over time, is modeled separately using a pair copula decomposition. Secondly we select a multivariate copula to describe the relations of the responses. Then we extend the model by also supposing that the dependence structure across the variables changes over time. In this approach we select the copula to capture the dependence between the R response variables conditional to the past for each subject and then we model the serial dependence by applying a PCC (in relation to the time) to each conditional distribution of the responses. The result is a new flexible multivariate longitudinal model, which overcomes the problem of modeling simultaneous dependence between two or more non-normal responses over time. The paper is organized as follows. In Sect. 2 we describe the copula and the D vine copula. In Sect. 3 we present our proposal by supposing that the dependence between the responses does not change over time. Then we extend the model by relaxing that assumption. Finally, we illustrate the models by an example for two response variables.

## 2 Copulae and D-vine Copulae

A bivariate copula is a function $C : I^2 \to I$, with $I^2 = [0, 1] \times [0, 1]$ and $I = [0, 1]$, that, with an appropriate extension in $R^2$ of the domain, satisfies all the properties of a cumulative distribution function (cdf). In particular, it is the cdf of a bivariate random variable $(U, V)$, with uniform marginal distributions in [0, 1]:

$$C_\lambda(u, v) = P(U \leq u, V \leq v; \lambda), \quad 0 \leq u \leq 1 \quad 0 \leq v \leq 1$$

where $\lambda$ is a parameter measuring the dependence between $U$ and $V$. The following theorem by Sklar [5] explains the use of the copula in the characterization of a joint distribution. Let $(Y_1, Y_2)$ be a bivariate random variable with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cdf $F_{Y_1,Y_2}(y_1, y_2; \lambda)$, then a copula function always exists $C_\lambda(\cdot, \cdot)$ with $C : I^2 \to I$ such that

$$F_{Y_1,Y_2}(y_1, y_2; \lambda) = C_\lambda\big(F_{Y_1}(y_1), F_{Y_2}(y_2)\big), \quad y_1, y_2 \in \mathbb{R}. \tag{1}$$

If the marginal cdfs are continuous functions then the copula $C(\cdot, \cdot)$ is unique. Moreover, if $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous the copula can be found inverting (1):

$$C_\lambda(u, v) = F_{Y_1,Y_2}(F_{Y_1}^{-1}(u), F_{Y_2}^{-1}(v)), \tag{2}$$

with $u = F_{Y_1}(y_1)$ and $v = F_{Y_2}(y_2)$. This theorem states that each joint cdf can be expressed in terms of two separate but related issues: the marginal distributions and the dependence structures between them. The dependence structure is described by the copula. Equation (1) provides a general mechanism to construct new bivariate cdfs in a straightforward manner. Since in high dimension the multivariate copulae usually used in the literature suffer from rather inflexible structure, alternative copula based constructions of multivariate distributions have been suggested. In particular vine pair-copula constructions (PCCs) have proven to be popular for the flexible modelling of multivariate dependence in numerous situations. Important work includes [1–4], while a recent overview was provided by [7]. In vine copulae the multivariate density function is decomposed as the product of bivariate copula densities (pair-copulae) on $[0, 1]^2$ and the marginal density functions. In d-dimension let $(Y_1, \ldots, Y_d)'$ be a random vector with joint cdf $F$ and df $f$, respectively. Consider the decomposition:

$$f(y_1, \ldots, y_d) = f(y_d|y_{d-1}, \ldots, y_1)f(y_1, \ldots, y_{d-1}) = \prod_{t=2}^{d} f(y_t|y_{t-1}, \ldots, y_1)f(y_1), \tag{3}$$

where $f(\cdot|\cdot)$ denotes the conditional density. Using Sklar's theorem for conditional bivariate densities, $f(y_t|y_{t-1}, \ldots, y_1)$ becomes:

$$\begin{aligned} f(y_t|y_{t-1} \ldots, y_1) &= \frac{f(y_t, y_1|y_{t-1}, \ldots, y_2)}{f(y_t|y_{t-1}, \ldots, y_2)} \qquad (4) \\ &= c_{t,1|t-1,\ldots,2}(F(y_t|y_{t-1}, \ldots, y_2), F(y_1|y_{t-1}, \ldots, y_2))f(y_t|y_{t-1}, \ldots, y_2), \end{aligned}$$

where $F(\ldots|\ldots)$ denotes a conditional cdf.

We adopt a simplification: for arbitrary distinct indices $t, s$, with $t > s$ we use the following abbreviation for a bivariate conditional copula density of $Y_t$ and $Y_s$ given $t - 1, \ldots, s + 1$

$$c_{t,s|t-1,\ldots,s+1} := c_{t,s|t-1,\ldots,s+1}(F(y_t|y_{t-1}, \ldots, y_{s+1}), F(y_s|y_{t-1}, \ldots, y_{s+1})). \tag{5}$$

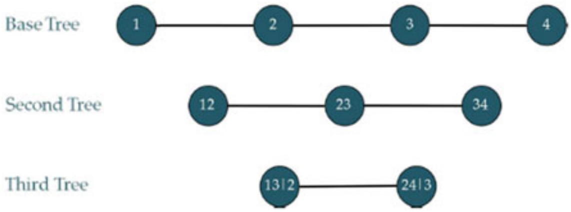


**Fig. 1** D-vine copula for 4 variables

Iteratively applying Sklar's theorem as in (4) and discarding arguments of conditional copulas, we obtain the following factorization for (3):

$$f(y_1,\ldots,y_d)=\prod_{t=2}^{d}\left\{\prod_{j=1}^{t-1}c_{t,j|t-1,\ldots,j+1}f(y_t)\right\}f(y_1), \tag{6}$$

which is a product of $d$ marginal densities and $d(d-1)/2$ pair-copula densities. The Eq. (6) can be recognised as D-vine model. D-vine copula is one of a wider class of graphical models discussed by [2]. Bedford and Cooke [2] arrange the pair-copula representation (6) using a sequence of nested trees with undirected edges, which they call a regular vine. Edges in the trees indicate the indices used for the conditional copula densities. Figure 1 shows the tree representation of a D-vine in four dimensions. It consists of trees arranged in three levels. An edge of a tree corresponds to a pair copula density denoted by the edge label. The whole structure is easy to construct and is helpful in understanding the corresponding PCC, that is:

$$\begin{aligned} c(F_1(y_1),\ldots,F_4(y_4)) &= c_{12}\ c_{23}\ c_{34} \\ &= c_{13|2}\ c_{24|3} \\ &= c_{14|23}. \end{aligned} \tag{7}$$

The D-vine is suitable for modeling the dependence in time series (see [6]).

## 3 Our Proposal

In this section we suggest a model for multivariate longitudinal data. First we suppose the existence of a dependence between the responses invariant over time. Therefore, the change over time of the distribution of the responses is due only to a change of the marginal conditional (to the past) distributions of each responses and not to the change of the dependence structure across the responses. The proposed model considers two different levels of analysis. At first, a multivariate copula describes the relations of the responses observed at a specific time. Then each longitudinal series, corresponding to a given response over time, is modeled separately using a pair copula decomposition to relate the distributions of the variables describing the observation given in different times. Then we extend the model by also supposing that the dependence structure across the variables changes over time. Let $C_{\boldsymbol{\lambda}}$ be a multivariate copula with parameter $\boldsymbol{\lambda}$ of the multivariate response variable

$$(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \dots, \mathbf{Y}^{(R)})$$

such that the joint cumulative distribution function (cdf) is

$$F(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(R)}) = C_{\boldsymbol{\lambda}}(F_1(\mathbf{y}^{(1)}), F_2(\mathbf{y}^{(2)}), \dots, F_R(\mathbf{y}^{(R)})), \tag{8}$$

and the joint density function (df) is

$$f(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(R)}) = c_{\boldsymbol{\lambda}}(F_1(\mathbf{y}^{(1)}), F_2(\mathbf{y}^{(2)}), \dots, F_R(\mathbf{y}^{(R)})) \prod_{r=1}^{R} f_r(\mathbf{y}_r). \tag{9}$$

Since we consider longitudinal data each response is observed over time on a sample of n subject. We model each continuous series which generates the longitudinal data using a pair copula decomposition (as in [6]). In this way we decompose the distribution of the process at certain point in time, conditional to the past, into the product of a sequence of bivariate copula density and marginal density. The advantage is that the marginal distribution of the process at each point can be modeled arbitrarily, while the dependence over time is captured by a sequence of bivariate copulae.

Let $(y_1, y_2, \dots, y_T)$ be the univariate series for the $r$-th response variable (we adopt a simplification of the notation dropping the index $r$), the joint density function can be decomposed in a product of the conditional (to the past) distributions:

$$f(y_1, y_2, \dots, y_T) = \prod_{i=1}^{T} f(y_t | y_{t-1}, \dots, y_1) f(y_1).$$

By using a pair copula decomposition we have:

$$f(y_t | y_{t-1}, \dots, y_1) = \prod_{j=1}^{t-1} c_{t,j}(F(y_t | y_{t-1}, \dots, y_{j+1}), F(y_j | y_{t-1}, \dots, y_{j+1}); \theta_{t,j}) f(y_t),$$

where $F(y_t)$ and $f(y_t)$ are the cdf and the df of the marginal $Y_t$ and $c_{t,j} = c_{t,j|t-1,t-2,\dots,j+1}$ are the pair copulae with parameters $\theta_{t,j}$. Therefore, the joint distribution of the process becomes a D-vine copula model of order $T$, which is a product of $T$ marginal densities and $T(T-1)/2$ pair copula densities:

$$f(y_1, y_2, \dots, y_T) = \prod_{t=2}^{T} \left[ \prod_{j=1}^{t-1} c_{t,j}(u_{t|j+1}, u_{j|t-1}; \theta_{t,s}) f(y_t) \right] f(y_1), \tag{10}$$

where $u_{t|j+1} = F(y_t|y_{t-1}, \dots, y_{j+1})$ and $u_{j|t-1} = F(y_j|y_{t-1}, \dots, y_{j+1})$.

By substituting (10) and the correspondent cumulative distribution function in (9) we obtain the model for multivariate longitudinal data:

$$f(Y^{(1)}, Y^{(2)}, .., Y^{(R)}) = \prod_{r=1}^{R} \left( \prod_{t=2}^{T} \left[ \prod_{j=1}^{t-1} c_{t,j} \left( u^{(r)}_{t|j+1}, u^{(r)}_{j|t-1}; \theta^{(r)}_{t,s} \right) f(y^{(r)}_t) \right] f\left( y^{(r)}_1 \right) \right) \cdot$$

$$c_{\boldsymbol{\lambda}} \left( \prod_{t=2}^{T} \left[ \prod_{j=1}^{t-1} C^{(1)}_{t,j} (u^{(1)}_{t|j+1}, u^{(1)}_{j|t-1}; \theta^{(1)}_{t,s}) \right] \cdots, \prod_{t=2}^{T} \left[ \prod_{j=1}^{t-1} C^{(R)}_{t,j} (u^{(R)}_{t|j+1}, u^{(R)}_{j|t-1}; \theta^{(R)}_{t,s}) \right] \right) \tag{11}$$

In (11) $\lambda$ describes the dependence between the responses and $\theta^{(r)}_{i,j}$ describes the dependence between the $r$-th response at time $t$ and the one at time $j$. Finally, the model can be extended by also supposing that the dependence structure across the variables changes over time. We define the distribution of the $R$ response variables at time $t$ conditional to their past. Let now $y^{*(r)}_t = y^{(r)}_t | y^{(r)}_{t-1}, \dots, y^{(r)}_1$, we consider the factorization:

$$f(y^{*(1)}_t, y^{*(2)}_t, \dots, y^{*(R)}_t) = \prod_{r=2}^{R} \left[ f_{\lambda_t}(y^{*(r)}_t | y^{*(r-1)}_t, \dots, y^{*(1)}_t) \right] f(y^{*(1)}_t)$$

Then by applying the PCC (in relation to the time) to each conditional distribution we define the joint distribution as the extension of the D-vine model in $R$ dimensions.

$$\begin{aligned} & f(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(R)}) = \\ = & \prod_{r=1}^{R} \left\{ \prod_{t=2}^{T} \left[ \prod_{j=1}^{t-1} c_{t,j}(\mathbf{u}_{t|j+1}, \mathbf{u}_{j|t-1}; \theta_{t,s}) f_{\boldsymbol{\lambda}_t}(y^{*(r)}_t | y^{*(r-1)}_t, \cdots, y^{*(1)}_t) \right] \right. \\ & \left. \cdot \ f_{\boldsymbol{\lambda}_1}(y^{*(r)}_1 | y^{*(r-1)}_1, \cdots, y^{*(1)}_1) \right\} \end{aligned} \tag{12}$$

where $u_{t|j+1} = F(y^{*(r)}_t | y^{*(r-1)}_t, \dots, y^{*(1)}_t)$. The df $f_{\lambda_t}$ can also be defined by a PCC (between the responses). In (12) $\lambda_t$ describes the dependence between the responses at time $t$ and $\theta^{(r)}_{i,j}$ describes the dependence between the $r$-th response at time $t$ and the one at time $j$. Note that in both the proposed models we suppose that the response at time $t$ is independent of the past of the other variables. To illustrate the approach we analyse a longitudinal data from the data set Diet[1] with dimensions $R = 2, T = 5$ and $n = 26$.

[1] The dataset Diet is available on request to authors.

**Table 1** Copula and $C_{t:1,2}$ estimate between the responses at time $t$

| Copula | Family | Par |
|---|---|---|
| $C_{1:1,2}$ | C | 0.734 |
| $C_{2:1,2}$ | N | 0.253 |
| $C_{3:1,2}$ | F | 0.837 |
| $C_{4:1,2}$ | J | 1.189 |
| $C_{5:1,2}$ | F | −0.802 |

*Example 1* In the data set Diet two response variables (Weight and Trigliceridies) are observed on a sample of n = 26 subjects during $T = 5$ years. A direct association between the variables is possible. We apply the model described in Eq. (12). We estimate the model by the ML methods. The code of the algorithm is based on functions in the *R* packages CDVine and VineCopula.

A bivariate copula models is used previously to model the dependence between the responses. Table 1 show the Ml estimate of the parameter $\lambda_{t:1,2}$ of the copula with the best fit between the responses at time $t$. The copula is chosen among the principal bivariate copulae implemented in the packages.

Two D-vine copulae are applied to model the serial dependence of the two response variables. In particular Fig. 2 shows the two D-vine trees that represent the PCC of the conditional df $f(Y^{(2)}|Y^{(1)})$ and the one of the df $f(Y^{(1)})$, respectively. The order of the variables of the first tree follows the temporal order. The squares represent the nodes, while the grey lines represent the arcs. The names of the nodes can be read in the squares. The pair copula families are identified by the labels of the edges in the considered trees and the values corresponding to pair copula parameters $\theta_{t,j}$ can be read in the edge labels. The thicker the grey line the higher the dependence between the variables represented by the nodes.

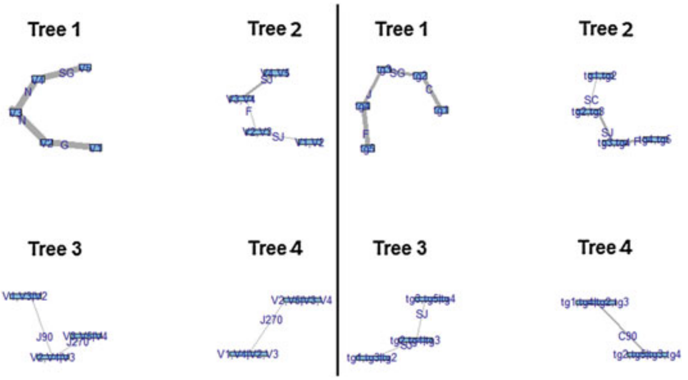


**Fig. 2** D-vine PCC trees for $f(y^{(2)}|y^{(1)})$ and $f(y^{(1)})$, respectively

## 4 Conclusion

This paper suggests the employment of PCC to model multivariate longitudinal data for capturing two types of dependence: between response variables at a given time and over time. The use of the PCC allows the description of the complex pattern of dependence of the multivariate longitudinal data and permits the construction of a flexible high-dimensional model by using only bivariate copulae as building blocks. The result is an extremly flexible model for multivariate longitudinal data, which overcomes the problem of modeling simultaneous dependence between two or more non-normal responses over time. However, the proposal can be extended by relaxing the assumption that the response variables are independent of the past of the other variables. Finally, a possible extension of this paper could be to extend the model to longitudinal data in which the length $T$ of the time series is not fixed but it varies from subject to subject. This is a typical problem in medicine or in clinical trials where the data are observed on patients over time. For example in cohort studies some patients drop out or new patients enter in the study during the period of experiment.

## References

1. Aas K, Czado C, Frigessi A, Bakken H (2009) Pair-copula constructions of multiple dependence. Insur Math Econ 182:198–244
2. Bedford T, Cooke RM (2002) Vines-a new graphical model for dependent random variables. Ann Stat 30:1031–1068
3. Haff I, Aas K, Frigessi A (2010) On the simplified pair-copula construction-Simply useful or too simplistic? J Multivariate Anal 101:1296–1310
4. Min A, Czado C (2010) Bayesian inference for multivariate copulas using pair-copula constructions. J Finan Econ 8:511–546
5. Nelsen RB (2013) An introduction to copulas. Springer, Science Business Media
6. Smith M, Min A, Almeida C, Czado C (2012) Modeling longitudinal data using a pair-copula decomposition of serial dependence. J Am Stat Assoc 105:1467–1479
7. Smith M (2015) Copula modelling of dependence in multivariate time series. Int J Forecast 31:815–833
8. Wolfinger RD (1993) Covariance structure selection in general mixed models. Commun Stat Simul Comput 22:1079–1106
9. Zeger SL, Liano KY, Self SG (1985) The analysis of binary longitudinal data with time independent covariates. Biometrika 72:31–38

# Predictability in Probabilistic Discrete Event Systems

**Farid Nouioua, Philippe Dague and Lina Ye**

**Abstract** Predictability is a key property allowing one to expect in advance the occurrence of a fault in a system based on its observed events. Existing works give a binary answer to the question of knowing whether a system is predictable or not. In this paper, we consider discrete event systems where probabilities of the transitions are available. We show how to take advantage of this information to perform a Markov chain-based analysis and extract probability values that give a finer appreciation of the degree of predictability. This analysis is particularly important in case of non predictable systems.

## 1 Introduction

Faults diagnosability is a key property to increase the autonomy of nowadays systems. This property has been extensively studied in the last years. The seminal work in [9] provided an algorithm to verify diagnosability in discrete event systems (DES) represented by finite automata, based on the so-called deterministic diagnoser. Subsequent works proposed polynomial algorithms, based on the twin plant approach [6, 12]. Diagnosability ensures the ability to detect faults after their occurrences. However, since it is not always easy to recover the system after the faults occurred, a stronger property has to be considered: the ability to predict the faults before their occurrences. This is very useful in practice since when the fault is predicted, appropriate measures may be taken to avoid its negative effects. In [4] the diagnoser and the twin plant approaches have been adapted to verify predictability. The work in

F. Nouioua (✉)
LSIS, CNRS, University Aix-Marseille, Marseille, France
e-mail: farid.nouioua@lsis.org

P. Dague
LRI, CNRS, University Paris-Sud, University Paris-Saclay, Orsay, France
e-mail: philippe.dague@lri.fr

L. Ye
LRI, CentraleSupélec, CNRS, University Paris-Saclay, Orsay, France
e-mail: lina.ye@lri.fr

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_47

[5] concerns the predictability of patterns and that in [2] deals with timed DES. The predictability of distributed DES has been studied in [11].

In the previous works, the decision about predictability tells simply either the system is predictable or not. However, if a system contains only a low proportion of traces where the fault cannot be predicted while a second one contains a much greater proportion of such traces, it would be plausible to associate a measure of non predictability that is more important in the latter system than in the former one. This kind of measure may be beneficial in practice. For instance, it may be better in some contexts to tolerate a system with a sufficiently low degree of non-predictability than to add the missing sensors to ensure predictability which can be very expensive. The work in [3] considers stochastic DESs and provides necessary and sufficient conditions for the notion of AAS-predictability (asymptotically almost sure predictability) which is the counterpart of the notion of AA-Diagnosability introduced in [10]. In [1], the authors propose different variants of predictability and diagnosability in stochastic DESs and show that checking diagnosability in this setting is PSPACE-Complete while checking predictability is NLOGSPACE-Complete. However the optimal size of both the diagnoser and the predictor remains exponential.

The present paper extends the approach proposed in [8] for diagnosability of probabilistic DES to deal with predictability. We propose a so-called light estimator which is a probabilistic DES allowing us to analyze predictability by extracting an appropriate Markov chain that explains the dynamics of the system. The results of the asymptotic behavior of this chain determine probability values that help one in quantifying the degree of non predictability.

The paper is organized as follows. The probabilistic model is presented in Sect. 2. Section 3 recalls the diagnoser-based approach for predictability. Section 4 presents the light estimator. Section 5 is devoted to the probabilistic analysis. Finally, Sect. 6 concludes the paper.

## 2 Probabilistic Discrete Event Model

The model used in this paper is that of a probabilistic DES (PDES) which consists of a classical DES enriched by probability values on its transitions.

**Definition 1** A *(PDES)* is modeled by the structure $\Gamma = (X, E, \theta, x_0)$ where $X = \{x_0, ..., x_{n-1}\}$ is a finite set of states ($|X| = n$), $E = \{e_0, ..., e_{m-1}\}$ is a finite set of events ($|E| = m$), $x_0$ is the initial state and $\theta : X \times E \times X \longrightarrow [0, 1]$ is a probabilistic transition function: $\theta(x, e, x') = \alpha$ $(0 \leq \alpha \leq 1)$ is the probability that the event $e$ occurs in $x$ and causes the transition of the system from state $x$ to state $x'$. We represent in the model all the possible transitions of the system in each state. Thus, for each $x \in X$: $\sum_{y \in X} \sum_{e \in E} \theta(x, e, y) = 1$.

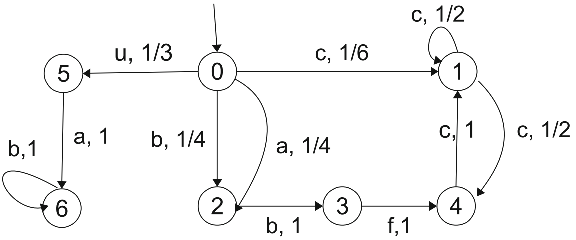


**Fig. 1** **a** Example of a probabilistic DES

To a PDES $\Gamma = (X, E, \theta, x_0)$ we associate a classical DES $G = (X, E, \delta, x_0)$ where the transition function $\delta : X \times E \times X \longrightarrow \{0, 1\}$ is defined by: $\delta(x, e, x') = 1$ if $\theta(x, e, x') > 0$ and $\delta(x, e, x') = 0$ otherwise.

$E^*$ denotes the Kleene closure of $E$. $\delta$ extends to words $s \in E^*$ with $s = a_1 \dots a_k$ by: $(x, s, x') \in \delta$ iff there is sequence of states $x_{j_0}, \dots, x_{j_k}$ such that $x_{j_0} = x$, $x_{j_k} = x'$ and $(x_{j_{i-1}}, a_i, x_{j_i}) \in \delta$ for $1 \leq i \leq k$.

We denote by $L(G) \subseteq E^*$ (or $L$ when no ambiguity), the language generated by $G$. $L(G)$ is prefix closed. The set of events $E$ is such that $E = E_o \cup E_{uo}$ where $E_o$ (resp. $E_{uo}$) contains the observable (resp. unobservable) events. $E_f \subseteq E_{uo}$ is the subset of unobservable faulty events. Moreover, faults are partitioned into disjoint sets corresponding to the different fault types: $E_f = E_{f_1} \cup \dots \cup E_{f_p}$. In the sequel, we will focus, without loss of generality, on one fault type as in [12]. For the sake of simplicity, we will denote by $f$ each occurrence of the considered fault type. We suppose also that $L(G)$ is live (there is at least one transition from any state in the system) and that there is no cycle in $G$ with only unobservable events.

*Example 1* Figure 1 shows a PDES $\Gamma = (X, E, \theta, x_0)$ where: $X = \{x_0, x_1, x_2, x_3, x_4, x_5, x_6\}$, $E_o = \{a, b, c\}$, $E_{uo} = \{f, u\}$ and $E_f = \{f\}$, the initial state is $x_0$ and the transition function is shown in Fig. 1.[1]

A word of $L(G)$ is also called trace. The empty trace is denoted by $\epsilon$. The post-language of $L(G)$ after a trace $s$ is: $L/s = \{t \in E^* | st \in L\}$. The set of prefixes of a word $s$ is denoted by $\overline{s}$.

$P : E^* \longrightarrow E_o^*$ is a projection function that erases from any trace its unobservable events: $P(\sigma) = \epsilon$ if $\sigma = \epsilon$ or $\sigma \in E_{uo}$, $P(\sigma) = \sigma$ if $\sigma \in E_o$ and $P(s\sigma) = P(s)P(\sigma)$ for $s \in E^*$ and $\sigma \in E$. $P_L^{-1}$ is the inverse projection: for any $w \in E_o^*$, $P_L^{-1}(w) = \{s \in L | P(s) = w\}$. It provides, for a sequence of observable events $w$, all traces of $L(G)$ whose projection is $w$.

$s_f$ denotes the final event of a trace $s$ and $\Psi(f)$ all traces ending in the fault event $f$: $\Psi(f) = \{s \in L | s_f = f\}$. We define: $X_o = \{x_0\} \cup \{x \in X \,|\exists y \in X, \exists e \in E_o, \delta(y, e, x) = 1\}$. $X_o$ includes the initial state $x_0$ and every state which is the target of at least one transition labelled by an observable event.

Let $L(G, x)$ denote the set of traces originating from $x$, $L_o(G, x)$ the subset of those traces of $L(G, x)$ that end at the first observable event and $L_\sigma(G, x)$ the subset

[1] To simplify figures, we represent a state name $x_i$ by its index $i$.

of $L_o(G,x)$ containing those traces that end at the observable event $\sigma$: $L_o(G,x) = \{s \in L(G,x) \mid s = u\sigma, u \in E_{uo}^*, \sigma \in E_o\}$, $L_\sigma(G,x) = \{s \in L_o(G,x) \mid s_f = \sigma\}$.

## 3 The "Binary" Predictability

Intuitively, a fault is predictable iff, based on observed events, one can deduce its occurrence, before it actually occurs.

**Definition 2** [4] $f$ is predictable iff: $(\exists n \in \mathbb{N})(\forall s \in \Psi(f))(\exists t \in \overline{s})[(f \notin t) \wedge P]$, where the predictability condition P is: $(\forall u \in L)(\forall v \in L/u)[(P(u) = P(t)) \wedge (f \notin u) \wedge (\|v\| \geq n) \Rightarrow (f \in v)]$.

Let us now recall the notion of diagnoser.

**Definition 3** A diagnoser is a deterministic automaton which is defined by $G_d = (Q_d, E_o, \delta_d, q_0)$ where:

- $Q_d \subseteq 2^{X_o \times \{N,F\}}$. A state of $Q_d$ is of the form: $q_d = \{(x_1, l_1), \ldots, (x_k, l_k)\}$ where $x_i \in X_o$, $l_i \in \{N, F\}$; $q_0 = \{(x_0, N)\}$ is the initial state of $G_d$;
- $E_o$ is the set of the observable events and
- $\delta_d : Q_d \times E_o \longrightarrow Q_d$ is the transition function of the diagnoser defined by: $\delta_d(q,\sigma) = \bigcup_{(x,l)\in q} \bigcup_{s \in L_\sigma(G,x)} \bigcup_{(x,s,x')\in\delta} \{(x', LP(l,s))\}$ where $LP : \{N,F\} \times E^* \longrightarrow \{N,F\}$ is a label propagation function defined by: if $l = N$ and $f \notin s$ then $LP(l,s) = N$ else $LP(l,s) = F$.

A state $q$ of $G_d$ is $f$-uncertain if $\exists (x,l), (x',l') \in q$ s.t. $l = N$ and $l' = F$. It is $f$-certain (resp. normal) if $\forall (x,l) \in q, l = F$ (resp. $\forall (x,l) \in q, l = N$).

We denote by $Q^N$ the set of normal states of $G_d$. Let $\mathcal{C}$ be the set of normal states having an immediate successor that is not normal. We call these states, the critical states: $\mathcal{C} = \{q \in Q^N \mid \exists q' = \delta_d(q,o)$ such that $o \in E_o$ and $q' \notin Q^N\}$. We put $\mathcal{C}_{OK} = \{q \in \mathcal{C} \mid$ all the accessible cycles from $q$ contain only $f$-certain states$\}$ and $\mathcal{C}_{KO} = \mathcal{C} \setminus \mathcal{C}_{OK}$. Then:

$f$ is predictable iff $\mathcal{C}_{KO} = \emptyset$:, i.e., every accessible cycle from any state of $\mathcal{C}$ is formed exclusively by $f$-certain states [4].

*Example 1* (*Cont*) Figure 2a depicts the diagnoser of the system presented in Example 1. We have: $\mathcal{C} = \{\{(3,N)\}, \{(3,N),(6,N)\}\}$. While the only accessible cycle from $\{3,N\}$ is constituted from $f$-certain states, we may reach from $\{(3,N),(6,N)\}$ either a cycle of normal states or a cycle of $f$-certain states. The fault $f$ is then not predictable.

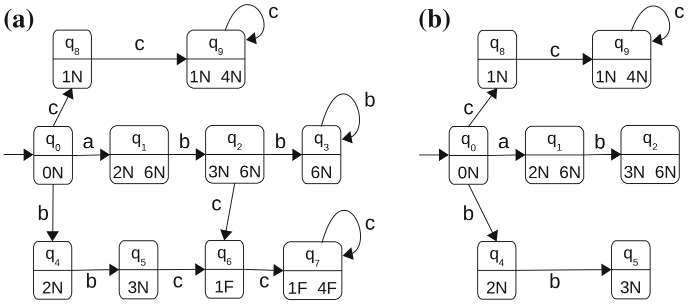


**Fig. 2** **a** Diagnoser $G_d$ **b** Simplified diagnoser $G'_d$

## 4 The Light Estimator

The light estimator (L-estimator) is constructed from a part of the diagnoser (henceforth called the simplified diagnoser) as follows:

**Definition 4** Let $G_d = (Q_d, E_o, \delta_d, q_0)$ be a diagnoser. The simplified diagnoser is defined by $G'_d = (Q'_d, E_o, \delta'_d, q_0)$ where $Q'_d = Q_d \setminus B$ where $B = \{q \in Q_d$: $q$ cannot be accessible from $q_0$ without passing by a state in $\mathcal{C}$ different from $q\}$ and $\delta'_d$ is the restriction of $\delta_d$ to $Q'_d$.

$G'_d$ is obtained from $G_d$ by: (1) removing all the transitions originating from any state $q \in \mathcal{C}$ then, (2) keeping only the part of the diagnoser that is accessible from the initial state $q_0$. Note that a state which is reachable both by passing by a state in $\mathcal{C}$ and without passing by such a state is kept, but only the paths from $q_0$ to that state that do not contain any state from $\mathcal{C}$ are kept. Note also that critical states no longer have outgoing transitions. Figure 2b depicts the simplified diagnoser of the system of Example 1.

The light estimator represents explicitly the transitions between the sub-states of the simplified diagnoser as well as their probability values.

**Definition 5** The L-estimator is defined by $H = (T, E'_o, \psi, t_0)$ where:

- Let $q_0, \ldots, q_m$ be the states of the simplified diagnoser $G'_d$ such that $q_0 = \{(x_0, N)\}$. The state space of $H$ is: $T \subseteq X_o \times \{N, F\} \times \{0, \ldots, m\}$ such that $(x, l, i) \in T$ iff $(x, l) \in q_i$;
- $t_0 = (x_0, N, 0)$ is the initial state;
- $E'_o = E_o \cup \{\alpha\}$ where $\alpha \notin E_o$ is a new event standing for any observable event. This event is added to ensure a coherent definition of the L-estimator and will not play any role in the following development;
- $\psi : T \times E'_o \times T \longrightarrow [0, 1]$ is the probabilistic transition function of $H$. Let $t = (x, l, i) \in T$ such that $q_i \notin \mathcal{C}$, $t' = (x', l', i') \in T$ and $\sigma \in E_o$. $\psi(t, \sigma, t') \neq 0$ if and only if there is a possible transition from $t$ to $t'$. This corresponds to the case where $(q_i, \sigma, q_{i'}) \in \delta'_d$ and there is at least some trace $s \in L_\sigma(G, x)$ such that $l' = LP(l, s)$

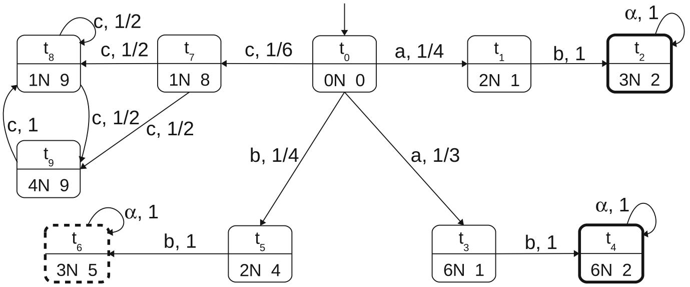


**Fig. 3** The light estimator

and $(x, s, x') \in \delta$. Let S be the set of all such traces: $S = \{s \in L_\sigma(G, x) | l' = LP(l, s)$ and $(x, s, x') \in \delta\}$. Then, $\psi(t, \sigma, t')$ is the sum of the probabilities of transitions from $x$ to $x'$ by the different traces of S: $\psi(t, \sigma, t') = \sum_{s \in S} \theta(x, s, x')$. For each state $t = (x, l, i)$ of $T$ such that $q_i \in \mathcal{C}$, we put $\psi(t, \alpha, t) = 1$.

Intuitively, a state $t = (x, l, i)$ of $H$ contains the relevant information about predictability. A trace $w$ of observable events leading from $t_0$ to $t = (x, l, i)$ in $H$ leads in the simplified diagnoser from $q_0$ to $q_i$ s.t. $(x, l) \in q_i$. Moreover, in case $q_i \in \mathcal{C}_{KO}$ (resp. $q_i \in \mathcal{C}_{OK}$), the fault may or may not occur (resp. will necessarily occur) before observing the next observable event after $w$ and in any observed trace having $w$ as prefix, the fault cannot be predicted. Thus, as soon as a state $t = (x, l, i)$ such that $q_i \in \mathcal{C}$ is reached, the decision about predictability can be taken independently from the subsequent continuations. This explains the addition of the loop on each such state with the probability 1. Figure 3 shows the L-estimator of the system given in Example 1 (states from $\mathcal{C}_{KO}$ (resp. $\mathcal{C}_{OK}$) are in bold (resp. dashed) line). Note that the sum of the probabilities of all transitions issued from each state of $H$ is 1.

## 5 Probabilistic Analysis

In this section, we show how to extract from the light estimator an homogeneous and discrete Markov chain and then to exploit the well known results about the asymptotic behaviors of such chains (for more details, see for example [7]) to obtain a finer appreciation of predictability. We believe that such a refinement is very useful in practice to deal with non-predictable systems.

To the light estimator $H = (T, E'_o, \psi, t_0)$, we associate the homogeneous and discrete time Markov chain $\{M_i, i = 0, 1...|T| - 1\}$ where $M_i$ is a random variable whose value is the state of the system after the observation of a set of events. $T$ is the state space of the Markov chain. The transition matrix $tr$ of the L-estimator is defined by: $\forall (t_1, t_2) \in T \times T$, $tr_{t_1,t_2} = \sum_{\sigma \in E'_o} \psi(t_1, \sigma, t_2)$.

*Example 1* (*Cont*) Since from each couple of states $(t, t')$ of $H$ there is at most one transition from $t$ to $t'$, the graphical representation of the Markov chain $\{M_i\}$ is obtained from Fig. 3 by just removing the observable events.

Now, from the study of the asymptotic behavior of the obtained Markov chain, we can compute relevant probability values for classes of possibly infinite observed traces of the system. This study follows the following steps:

- Classify the states of $\{M_i\}$. A class is a strongly connected component in the graph of $\{M_i\}$; a class is persistent if each of its states has no successor outside it, otherwise, it is transitory. Let $\zeta = \{C_1, \ldots, C_h\}$ (resp. $\mu = \{\mu_1, \ldots, \mu_r\}$) be the set of persistent (resp. transitory) classes.
- Put the transition matrix in the canonical form in which persistent classes are put at the beginning. We obtain:

$$tr = \begin{pmatrix} Tr_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & Tr_h & 0 \\ R_1 & \cdots & R_h & Q \end{pmatrix}$$

  $Tr_i$ is the matrix containing the transition probabilities inside the persistent class $C_i$. $R = [R_1, \ldots, R_h]$ (resp. $Q$) contains the transition probabilities from transitory to persistent (resp. to transitory) states.
- Compute the fundamental matrix $N = (I - Q)^{-1}$ (I is the unit matrix of size: $|\mu_1| + \cdots + |\mu_r|$) and the absorption matrix $B = N.R$. We have the following results: the probability to be in a transitory state after an infinite number of steps is 0; the probability of absorption in the persistent state $j$ when we start from state $i$ is $B_{i,j}$. The absorption probability of a persistent class is then the sum of the absorption probabilities of its states.

We suppose without loss of generality that the initial state $t_0$ is the first transitory state. Thus, we are interested only in the first rows of $N$ and $B$.

The last step is to extract relevant probabilities from the Markov chain. Let $\{M_i\}$ be the Markov chain associated to a L-estimator and let $N$ and $B$ its fundamental and absorption matrices respectively. Let $\mathcal{T}_{ko}$ (resp. $\mathcal{T}_{ok}$) be the subset of persistent classes whose states correspond to critical states of the diagnoser where the fault is not predictable (resp. predictable), i.e. states $t = (x, l, i)$ of the estimator where $q_i \in \mathcal{C}_{KO}$ (resp. where $q_i \in \mathcal{C}_{OK}$). Let $\mathcal{T}_{nf}$ be the subset of all the other persistent classes, i.e. where the fault does not occur.

Then, we can define the following relevant probabilities: $\mathcal{P}_{ko}$ (resp. $\mathcal{P}_{ok}$) is the probability to follow a trace where the fault cannot be predicted (resp. surely occurs and is predicted): $\mathcal{P}_{ko} = \sum_{c \in \mathcal{T}_{ko}} \sum_{t \in c} (B)_{0,t}$ (resp. $\mathcal{P}_{ok} = \sum_{c \in \mathcal{T}_{ok}} \sum_{t \in c} (B)_{0,t}$). $\mathcal{P}_{nf}$ is the probability to follow a trace where the fault surely does not occur. $\mathcal{P}_{nf} = \sum_{c \in \mathcal{T}_{nf}} \sum_{t \in c} (B)_{0,t} = 1 - (\mathcal{P}_{ko} + \mathcal{P}_{ok})$.

*Example 1* (*Cont*) In our example, the persistent classes are: $C_1 = \{t_2\}$, $C_2 = \{t_4\}$, $C_3 = \{t_6\}$ and $C_4 = \{t_8, t_9\}$ where $\mathcal{T}_{ko} = \{C_1, C_2\}$, $\mathcal{T}_{ok} = \{C_3\}$ and $\mathcal{T}_{nf} = \{C_4\}$. The first rows of the matrices $N$ and $B$ are:

$$N_0 = \begin{array}{c} \begin{array}{ccccc} t_0 & t_1 & t_3 & t_5 & t_7 \end{array} \\ \left(\begin{array}{ccccc} 1 & \frac{1}{4} & \frac{1}{3} & \frac{1}{4} & \frac{1}{6} \end{array}\right) \end{array} \quad \text{and} \quad B_0 = \begin{array}{c} \begin{array}{ccccc} t_2 & t_4 & t_6 & t_8 & t_9 \end{array} \\ \left(\begin{array}{ccccc} \frac{1}{4} & \frac{1}{3} & \frac{1}{4} & \frac{1}{12} & \frac{1}{12} \end{array}\right) \end{array}$$

We obtain the values: $\mathcal{P}_{ko} = 7/12$, $\mathcal{P}_{ok} = 1/4$ and $\mathcal{P}_{nf} = 1/6$. This means that we have a probability of 7/12 (resp. 1/4) to be in a trace where the fault cannot be predicted (resp. the fault will occur and it is predicted) and a probability of 1/6 to be in a trace where the fault will not occur.

## 6 Conclusion

This paper investigated the use of information about probabilities of transitions in a DES to refine the decision about fault predictability. In particular, the proposed approach allows one to quantify the degree of non-predictability and accordingly to deal in a more flexible way with non predictable systems.

We plan also to generalize the probabilistic-based approach to the pattern predictability, to distributed systems and to other DESs such as Petri nets.

## References

1. Bertrand N, Haddad S, Lefaucheux E (2014) In: Foundation of Diagnosis and Predictability in Probabilistic Systems. IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science
2. Cassez F, Grastien A (2013) Predictability of event occurrences in timed systems. In: International workshop on formal modeling and analysis of timed systems. pp 417–429
3. Chang M, Dong W, Ji Y, Tong L (2013) On fault predictability in stochastic discrete event systems. Asian J Control 15(5):1458–1467
4. Genc S, Lafortune S (2009) Predictability of event occurrences in partially-observed discrete-event systems. Automatica 45(2):301–311
5. Jeron T, Marchand H, Genc S, Lafortune S (2008) Predictability of sequence patterns in discrete event systems. In: IFAC world congress. pp 537–543
6. Jiang S, Huang Z, Chandra V, Kumar R (2001) A polynomial algorithm for testing diagnosability of discrete event systems. IEEE Trans Autom Control 46(8):1318–1321
7. Kemeny JG, Snell JL (1976) Finite markov chains. Springer
8. Nouioua F, Dague P (2008) A probabilistic analysis of diagnosability in discrete event systems. In: European conference on artificial intelligent. pp 224–228
9. Sampath M, Sengupta R, Lafortune S, Sinnamohideen K, Teneketzis D (1995) Diagnosability of discrete-event systems. IEEE Trans Autom Control 40(9):1555–1575
10. Thorsley D, Teneketzis D (2005) Diagnosability of stochastic discrete-event systems. IEEE Trans Autom Control 50(4):476–492

11. Ye L, Dague P, Nouioua F (2013) Predictability analysis of distributed discrete event systems. In: IEEE conference on decision and control. pp 5009–5015
12. Yoo T, Lafortune S (2002) Polynomial-time verification of diagnosability of partially-observed discrete-event systems. IEEE Trans Autom Control 47(9):1491–1495

# A Sandwich Theorem for Natural Extensions

**Renato Pelessoni and Paolo Vicig**

**Abstract** The recently introduced weak consistency notions of 2-coherence and 2-convexity are endowed with a concept of 2-coherent, respectively, 2-convex natural extension, whose properties parallel those of the natural extension for coherent lower previsions. We show that some of these extensions coincide in various common instances, thus producing the same inferences.

**Keywords** 2-convex lower previsions · Coherent lower previsions · Natural extensions

## 1 Introduction

In a recent paper [4] we introduced two weak consistency concepts for conditional lower previsions, 2-*convexity* and 2-*coherence*, studying their basic properties in greater detail in [5]. Formally, 2-coherent and 2-convex conditional lower previsions are a broad generalisation of the 2-coherent (unconditional) lower previsions in [6, Appendix B]. Our main aim in introducing them was to explore the flexibility of the betting scheme which underlies these and other consistency concepts (starting with de Finetti's subjective probability [1]), showing the capability of these previsions of encompassing a number of different uncertainty models in a unified framework.

An important issue is also to detect which properties from stronger consistency concepts are somehow retained by either 2-convexity or 2-coherence. As shown in [4, 5], a very relevant feature of theirs is that they are endowed with, respectively, a 2-convex and a 2-coherent *natural extension*. The properties of these extensions, exemplified in Proposition 1, are formally perfectly analogous to those of the natural extension for coherent lower previsions (following Williams' coherence in the conditional framework [7]) or the convex natural extension for convex conditional

R. Pelessoni (✉) · P. Vicig
DEAMS 'B. de Finetti', University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy
e-mail: renato.pelessoni@econ.units.it

P. Vicig
e-mail: paolo.vicig@econ.units.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_48

previsions [2]. In particular, when finite, they allow extending a lower prevision $\underline{P}$ from its domain $\mathcal{D}$ to any larger $\mathcal{D}' \supset \mathcal{D}$. Yet, when different natural extensions can be applied to the same $\underline{P}$, the results may differ also considerably (cf. the later Example 1 in Sect. 3). Since 2-coherence is weaker than coherence, inferences produced by the 2-coherent natural extension will be generally vaguer than those guaranteed by the coherent natural extension, and similarly with other instances. Actually, often 2-coherent or 2-convex natural extensions will be even too vague. This points out a drawback of these weak consistency notions and is one reason why, in our view, they should not be regarded as realistic candidates for replacing coherence or convexity. Rather, we will show in this paper that they may be helpful precisely for determining the coherent natural extension, or the convex natural extension. In fact, after concisely presenting the necessary preliminary notions in Sect. 2, we show in Sect. 3 that there are significant instances where some or all of the four extensions we mentioned so far coincide. For this, the lower prevision $\underline{P}$ is initially defined on a structured set $\mathcal{X}|\mathcal{B}^{\varnothing}$ (cf. Definition 2) of conditional gambles, representing a generalisation of a vector space to a conditional environment. Hence we are considering a special, but rather common, situation. In Proposition 2 we give an alternative expression for the coherent natural extension, which is later needed and generalises a result in [6] (cf. Corollary 1). After showing how to ensure finiteness for the relevant natural extensions, Theorems 2, 3 and 4 present instances where more different extensions coincide. These results are discussed in the comment after Theorem 4 and in the concluding Sect. 4. Due to space constraints, some of the proofs are omitted.

## 2 Preliminaries

Let $\mathcal{D}$ be an arbitrary set of conditional gambles, that is, the generic element of $\mathcal{D}$ is $X|B$, with $X$ a gamble (a bounded random variable), and $B$ non-impossible event. A *conditional lower prevision* $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ is a real map which, behaviourally, determines the supremum buying price $\underline{P}(X|B)$ of any $X|B \in \mathcal{D}$. This means that an agent should be willing to buy, or to bet in favour of, $X|B$, for any price lower than $\underline{P}(X|B)$. The agent's *gain* from the transaction/bet on $X|B$ for $\underline{P}(X|B)$ is $I_B(X - \underline{P}(X|B))$. Here $I_B$ is the indicator of event $B$. Its role is that of ensuring that the purchased bet is called off and the money returned to the agent iff $B$ does not occur. In the sequel, we shall use the symbol $B$ for both event $B$ and its indicator $I_B$.

A generic consistency requirement for $\underline{P}$ asks that no finite linear combination of bets on elements of $\mathcal{D}$, with prices given by $\underline{P}$, should produce a loss (bounded away from 0) for the agent. We obtain different known concepts by imposing constraints on the number of terms in the linear combination or on their coefficients $s_i$:

**Definition 1** Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be a given conditional lower prevision.

(a) $\underline{P}$ is a *coherent* conditional lower prevision on $\mathcal{D}$ iff, for all $m \in \mathbb{N}_0, \forall X_0|B_0, \ldots, X_m|B_m \in \mathcal{D}$, $\forall s_0, \ldots, s_m \geq 0$, defining $S(\underline{s}) = \bigvee\{B_i : s_i \neq 0, i = 0, \ldots, m\}$

and $\underline{G} = \sum_{i=1}^{m} s_i B_i (X_i - \underline{P}(X_i|B_i)) - s_0 B_0 (X_0 - \underline{P}(X_0|B_0))$, it holds, whenever $S(\underline{s}) \neq \varnothing$, that $\sup\{\underline{G}|S(\underline{s})\} \geq 0$.

(b) $\underline{P}$ is 2-*coherent* on $\mathcal{D}$ iff a) holds with $m = 1$ (hence there are *two* terms in $\underline{G}$).

(c) $\underline{P}$ is *convex* on $\mathcal{D}$ iff a) holds with the additional convexity constraint $\sum_{i=1}^{m} s_i = s_0 = 1$.

(d) $\underline{P}$ is 2-*convex* on $\mathcal{D}$ iff c) holds with $m = 1$, i.e., iff, $\forall X_0|B_0,\ X_1|B_1 \in \mathcal{D}$, we have that, defining $\underline{G}_{2c} = B_1(X_1 - \underline{P}(X_1|B_1)) - B_0(X_0 - \underline{P}(X_0|B_0))$, $\sup(\underline{G}_{2c}| B_0 \vee B_1) \geq 0$.

(e) $\underline{P}$ is *centered*, convex or 2-convex, on $\mathcal{D}$ iff it is convex or 2-convex, respectively, and $\forall X|B \in \mathcal{D}$, we have that $0|B \in \mathcal{D}$ and $\underline{P}(0|B) = 0$.

Condition a), which is Williams' coherence [7] in the structure-free version of [3], is clearly the strongest one. Convexity is a relaxation of coherence, studied in [2]. Given $\underline{P}$ on $\mathcal{D}$, the following relationships hold:

$$\begin{array}{l} \underline{P} \text{ coherent} \Rightarrow \underline{P} \text{ 2-coherent} \Rightarrow \underline{P} \text{ 2-convex} \\ \underline{P} \text{ coherent} \Rightarrow \underline{P} \text{ convex} \Rightarrow \underline{P} \text{ 2-convex.} \end{array} \tag{1}$$

The consistency concepts recalled so far can be characterised by means of axioms on the special sets $\mathcal{X}|\mathcal{B}^{\varnothing}$ defined next:

**Definition 2** Let $\mathcal{X}$ be a linear space of gambles and $\mathcal{B} \subset \mathcal{X}$ a set of (indicators of) events, such that $\Omega \in \mathcal{B}$ and $BX \in \mathcal{X}, \forall B \in \mathcal{B}, \forall X \in \mathcal{X}$. Setting $\mathcal{B}^{\varnothing} = \mathcal{B} - \{\varnothing\}$, define $\mathcal{X}|\mathcal{B}^{\varnothing} = \{X|B : X \in \mathcal{X},\ B \in \mathcal{B}^{\varnothing}\}$.

**Theorem 1** (Characterisation Theorems) *Let* $\underline{P} : \mathcal{X}|\mathcal{B}^{\varnothing} \to \mathbb{R}$ *be a conditional lower prevision.*

*(a)* $\underline{P}$ *is coherent on* $\mathcal{X}|\mathcal{B}^{\varnothing}$ *if and only if [3, 7]*

*(A1)* $\underline{P}(X|B) - \underline{P}(Y|B) \leq \sup\{X - Y|B\}, \forall X|B, Y|B \in \mathcal{X}|\mathcal{B}^{\varnothing}$.

*(A2)* $\underline{P}(\lambda X|B) = \lambda \underline{P}(X|B), \forall X|B \in \mathcal{X}|\mathcal{B}^{\varnothing}, \forall \lambda \geq 0$.

*(A3)* $\underline{P}(X + Y|B) \geq \underline{P}(X|B) + \underline{P}(Y|B), \forall X|B,\ Y|B \in \mathcal{X}|\mathcal{B}^{\varnothing}$.

*(A4)* $\underline{P}(A(X - \underline{P}(X|A \wedge B))|B) = 0, \forall X \in \mathcal{X}, \forall A,\ B \in \mathcal{B}^{\varnothing} : A \wedge B \neq \varnothing$.

*(b)* $\underline{P}$ *is* 2*-coherent on* $\mathcal{X}|\mathcal{B}^{\varnothing}$ *if and only if (A1), (A2), (A4) and the following axiom hold [5]:*

*(A5)* $\underline{P}(\lambda X|B) \leq \lambda \underline{P}(X|B), \forall \lambda < 0$.

*(c)* $\underline{P}$ *is convex on* $\mathcal{X}|\mathcal{B}^{\varnothing}$ *if and only if (A1), (A4) and the following axiom hold [2, Theorem 8]*

*(A6)* $\underline{P}(\lambda X + (1-\lambda)Y|B) \geq \lambda \underline{P}(X|B) + (1-\lambda)\underline{P}(Y|B), \forall X|B, Y|B \in \mathcal{X}|\mathcal{B}^{\varnothing}, \forall \lambda \in ]0, 1[$.

*(d)* $\underline{P}$ *is* 2*-convex on* $\mathcal{X}|\mathcal{B}^{\varnothing}$ *if and only if (A1) and (A4) hold [5].*

Next we recall the definitions of the various natural extensions studied in this paper. The term 'natural extension', without further qualifications, will denote the coherent natural extension in Definition 3, (a).

**Definition 3** (*Various natural extensions*) Let $\underline{P} : \mathcal{D} \to \mathbb{R}$ be a conditional lower prevision, and $Z|A$ a conditional gamble.

(a) Define $L(Z|A) = \{\alpha : \sup\{\sum_{i=1}^{m} s_i B_i(X_i - \underline{P}(X_i|B_i)) - A(Z - \alpha)|A \vee S(\underline{s})\} < 0,$ for some $X_i|B_i \in \mathcal{D}, s_i \geq 0, i = 1, \dots, m\}$, where $S(\underline{s}) = \vee_{i=1}^{m}\{B_i : s_i \neq 0\}$. Then, the *(coherent)* natural extension of $\underline{P}$ on $Z|A$ is $\underline{E}(Z|A) = \sup L(Z|A)$.
(b) Define $L_2(Z|A)$ putting $m = 1$ in $L(Z|A)$. The 2-*coherent* natural extension of $\underline{P}$ on $Z|A$ is $\underline{E}_2(Z|A) = \sup L_2(Z|A)$.
(c) Define $L_c(Z|A)$ from $L(Z|A)$, by adding the constraint $\sum_{i=1}^{m} s_i = 1$ in the 'for some' part. The *convex* natural extension of $\underline{P}$ on $Z|A$ is $\underline{E}_c(Z|A) = \sup L_c(Z|A)$.
(d) Define $L_{2c}(Z|A)$ putting $m = 1$ in $L_c(Z|A)$, i.e. $L_{2c}(Z|A) = \{\alpha : \sup\{B(X - \underline{P}(X|B)) - A(Z - \alpha)|A \vee B\} < 0,$ for some $X|B \in \mathcal{D}\}$. Then, the 2-*convex* natural extension $\underline{E}_{2c}$ of $\underline{P}$ on $Z|A$ is $\underline{E}_{2c} = \sup L_{2c}(Z|A)$.

The properties of these four natural extensions are analogous [2, 3, 5]. Here we state them for the 2-convex natural extension. For the properties of $\underline{E}$, $\underline{E}_2$, $\underline{E}_c$, replace $\underline{E}_{2c}$ and '2-convex' with, respectively, $\underline{E}$ and 'coherent', $\underline{E}_2$ and '2-coherent', $\underline{E}_c$ and 'convex'.

**Proposition 1** *Let $\underline{P} : \mathcal{D} \to \mathbb{R}$ a conditional lower prevision, with $\mathcal{D} \subset \mathcal{D}^*$. If $\underline{E}_{2c}$ is finite on $\mathcal{D}^*$, then*

*(a) $\underline{E}_{2c}(Z|A) \geq \underline{P}(Z|A)$, $\forall Z|A \in \mathcal{D}$.*
*(b) $\underline{E}_{2c}$ is 2-convex on $\mathcal{D}^*$.*
*(c) If $\underline{P}^*$ is 2-convex on $\mathcal{D}^*$ and $\underline{P}^*(Z|A) \geq \underline{P}(Z|A)$, $\forall Z|A \in \mathcal{D}$, then $\underline{P}^*(Z|A) \geq \underline{E}_{2c}(Z|A)$, $\forall Z|A \in \mathcal{D}^*$.*
*(d) $\underline{P}$ is 2-convex on $\mathcal{D}$ if and only if $\underline{E}_{2c} = \underline{P}$ on $\mathcal{D}$.*
*(e) If $\underline{P}$ is 2-convex on $\mathcal{D}$, then $\underline{E}_{2c}$ is its smallest 2-convex extension on $\mathcal{D}^*$.*

## 3 When Do Different Natural Extensions Coincide?

Given a lower prevision $\underline{P}$ on $\mathcal{D}$, its natural extensions $\underline{E}$, $\underline{E}_2$, $\underline{E}_c$, $\underline{E}_{2c}$ will generally be different, and ordered (when finite) as follows.

**Lemma 1** *Given $\underline{P} : \mathcal{D} \to \mathbb{R}$, it holds that*

$$\begin{aligned} \underline{E} \geq \underline{E}_2 \geq \underline{E}_{2c} \\ \underline{E} \geq \underline{E}_c \geq \underline{E}_{2c}. \end{aligned} \tag{2}$$

*Proof* It is easy to realise that (2) holds recalling (1), Definition 3 and Proposition 1. For instance, $\underline{E}_c \geq \underline{E}_{2c}$ because $\underline{E}_c$, being convex (Proposition 1, (b)), is also 2-convex (cf. (1)), but then $\underline{E}_c \geq \underline{E}_{2c}$ by Proposition 1, (e). □

It may also be the case that some among $\underline{E}$, $\underline{E}_2$, $\underline{E}_c$, $\underline{E}_{2c}$ are infinite. But even when being finite, they may differ considerably, as illustrated by the next simple example.

*Example 1* Let $\mathcal{D} = \{X\}$, where $X$ may only take the values 0 and 1. Assign $\underline{P}(X) \in (0, 1)$, which is clearly coherent, hence 2-convex, on $\mathcal{D}$. Its natural extension $\underline{E}$ on $\{2X\}$ is $\underline{E}(2X) = 2\underline{P}(X)$ by (A2), because $\underline{E}$ is coherent on $\{X, 2X\}$ and coincides with $\underline{P}$ on $X$. However, $\underline{E}_{2c}(2X) \geq \underline{P}(X)$ by (A1) and $\underline{E}_{2c}(2X) = \underline{P}(X) < \underline{E}(2X)$ is 2-convex. This can be checked directly using Definition 1, (d). (There are only two gains $\underline{G}_{2c}$ to inspect.)

On our way to establish when more natural extensions may coincide, we preliminarily tackle two issues: derive an alternative expression for the (coherent) natural extension, and discuss how to hedge possibly non-finite extensions. We assume throughout that the lower prevision $\underline{P}$ is initially assessed on some set $\mathcal{X}|\mathcal{B}^{\varnothing}$. As for the former issue, the following proposition holds.

**Proposition 2** *Let $\underline{P}$ be coherent on $\mathcal{X}|\mathcal{B}^{\varnothing}$. Then, defining*

$$\begin{aligned} L_1(Z|A) = \{\alpha : \sup\{BX - A(Z-\alpha)|A \vee B\} < 0, \\ \textit{for some } X \in \mathcal{X},\ B \in \mathcal{B},\ \textit{with } \underline{P}(X|B) = 0 \text{ if } B \neq \varnothing\}, \end{aligned} \tag{3}$$

*$L_1(Z|A) = L(Z|A)$ and the natural extension of $\underline{P}$ on $Z|A$ is*

$$\underline{E}(Z|A) = \sup L_1(Z|A). \tag{4}$$

*Proof* We prove that $L_1(Z|A) = L(Z|A)$, with $L_1(Z|A)$ defined in (3), $L(Z|A)$ in Definition 3 (a); taking their suprema gives then the thesis.

(i) $L_1(Z|A) \subset L(Z|A)$.
In fact, let $\alpha \in L_1(Z|A)$. Then $\sup\{BX - A(Z-\alpha)|A \vee B\} < 0$. If $B = \varnothing$, then $BX = 0$, $A \vee B = A$ in the supremum argument, and $\alpha \in L(Z|A)$ (case $S(\underline{s}) = \varnothing$). If $B \neq \varnothing$, then $\underline{P}(X|B) = 0$ and writing the supremum as sup $\{B(X - \underline{P}(X|B)) - A(Z-\alpha)|A \vee B\} < 0$ it appears that again $\alpha \in L(Z|A)$.

(ii) $L(Z|A) \subset L_1(Z|A)$.
Let now $\alpha \in L(Z|A)$ and, referring to the definition of $L(Z|A)$, $W = \sum_{i=1}^{m} s_i B_i (X_i - \underline{P}(X_i|B_i)) - A(Z-\alpha)$.
*If* $S(\underline{s}) = \varnothing$, then $\sup\{-A(Z-\alpha)|A\} < 0$ ensures that $\alpha \in L_1(Z|A)$ (case $B = \varnothing$).
*If* $S(\underline{s}) \neq \varnothing$, since $\underline{P}$ is coherent on $\mathcal{X}|\mathcal{B}^{\varnothing}$, we may apply (A2), (A3) and (A4) in Theorem 1 (a) to get

$$\begin{aligned} \underline{P}(\textstyle\sum_{i:s_i\neq 0} s_i B_i(X_i - \underline{P}(X_i|B_i))|S(\underline{s})) \geq \\ \textstyle\sum_{i:s_i\neq 0} s_i \underline{P}(B_i(X_i - \underline{P}(X_i|B_i))|S(\underline{s})) = 0. \end{aligned} \tag{5}$$

Define $Y = \sum_{i:s_i\neq 0} s_i B_i(X_i - \underline{P}(X_i|B_i))$. Since $\underline{P}(Y|S(\underline{s})) \geq 0$ by (5), we obtain

$$S(\underline{s})[Y - \underline{P}(Y|S(\underline{s}))] - A(Z-\alpha) \leq Y - A(Z-\alpha) = W$$

and hence

$$\begin{aligned}&\sup\{S(\underline{s})[Y - \underline{P}(Y|S(\underline{s}))] - A(Z - \alpha)|A \vee S(\underline{s})\} \leq \\ &\sup\{W|A \vee S(\underline{s})\} < 0.\end{aligned} \tag{6}$$

Now put $S(\underline{s}) = B, Y - \underline{P}(Y|B) = X$, and note that $\underline{P}(X|B) = \underline{P}(Y - \underline{P}(Y|B)|B) = \underline{P}(Y|B) - \underline{P}(Y|B) = 0$, recalling $\underline{P}(Y - c|B) = \underline{P}(Y|B) - c$, a necessary condition for coherence, at the second equality. Hence (6) may be rewritten as

$$\sup\{BX - A(Z - \alpha)|A \vee B\} < 0,$$

which proves that $\alpha \in L_1(Z|A)$.

□

While (4) supplies a new alternative expression for $\underline{E}(Z|A)$, it is interesting to observe that it boils down to a known result in the unconditional case, formally obtained putting $\mathcal{B} = \{\Omega, \emptyset\}$, $A = \Omega$ in Proposition 2.

**Corollary 1** *If $\underline{P}$ is coherent on a linear space $\mathcal{X}$, then*

$$\underline{E}(Z) = \sup\{\underline{P}(X) : X \leq Z, X \in \mathcal{X}\}. \tag{7}$$

In fact, Corollary 1 is part of the statement of Corollary 3.1.8 in [6].

Turning to the second issue, we are interested in guaranteeing that the various natural extensions considered are finite, i.e. neither $-\infty$ nor $+\infty$. Regarding $\underline{E}$ (or $\underline{E}_2$), its finiteness is ensured if the lower prevision $\underline{P}$ to be extended is coherent (or 2-coherent) [3, 5]. In the case of $\underline{E}_c$ or $\underline{E}_{2c}$, a sufficient condition [5] is that $\underline{P}(0|A) = 0$, for any additional $Z|A$ we wish to extend $\underline{P}$ to. While this condition is generally not necessary, it is nonetheless rather natural, but a 2-convex or convex $\underline{P}$ does not necessarily fulfil it. In fact, it may be the case that $0|A \in \mathcal{X}|\mathcal{B}^{\emptyset}$ and $\underline{P}(0|A) \neq 0$, which we can avoid by restricting our attention to *centered* 2-convex or convex previsions. But even doing so, *as we will*, it may happen that $0|A \notin \mathcal{X}|\mathcal{B}^{\emptyset}$ and, unlike the case of a coherent or 2-coherent $\underline{P}$, $\underline{P}(0|A) = 0$ is not the unique (2-)convex extension of $\underline{P}$. However, it holds that [2, 5]:

**Proposition 3** *Let $\underline{P}$ be centered 2-convex (alternatively, centered convex) on $\mathcal{X}|\mathcal{B}^{\emptyset}$. Given $0|A \notin \mathcal{X}|\mathcal{B}^{\emptyset}$, the extension of $\underline{P}$ such that $\underline{P}(0|A) = 0$ is 2-convex (convex).*

Proposition 3 suggests that when extending a centered $\underline{P}$ from $\mathcal{X}|\mathcal{B}^{\emptyset}$ to $\mathcal{D}^* \supset \mathcal{X}|\mathcal{B}^{\emptyset}$ we could consider first extending it to the set

$$(\mathcal{X}|\mathcal{B}^{\emptyset})^+ = \mathcal{X}|\mathcal{B}^{\emptyset} \cup \{0|A : Z|A \in \mathcal{D}^*\}, \tag{8}$$

putting $\underline{P}(0|A) = 0$. Adding zeroes is harmless when considering the natural extension, in the sense of the following

**Lemma 2** *Assign $\underline{P}$ on $\mathcal{X}|\mathcal{B}^{\varnothing}$ and let $\mathcal{D}^* \supset \mathcal{X}|\mathcal{B}^{\varnothing}$. Using the notation $L(Z|A)$ for the set $L$ in Definition 3 (a) when $\mathcal{D}$ there is replaced by $\mathcal{X}|\mathcal{B}^{\varnothing}$, we write $L^+(Z|A)$ instead when $\mathcal{D} = (\mathcal{X}|\mathcal{B}^{\varnothing})^+$. Then $L(Z|A) = L^+(Z|A)$, and consequently $\underline{E}(Z|A) = \sup L(Z|A) = \sup L^+(Z|A)$, $\forall Z|A \in \mathcal{D}^*$.*

**Definition 4** Given $\mathcal{X}|\mathcal{B}^{\varnothing} \subset \mathcal{D}^*$, let $\underline{P}$ be defined on $\mathcal{X}|\mathcal{B}^{\varnothing}$, and on $(\mathcal{X}|\mathcal{B}^{\varnothing})^+$ putting $\underline{P}(0|A) = 0, \forall 0|A \in (\mathcal{X}|\mathcal{B}^{\varnothing})^+$. Then, $\underline{E}_c^+$, $\underline{E}_{2c}^+$ are the convex, respectively 2-convex natural extension of $\underline{P}$ from $(\mathcal{X}|\mathcal{B}^{\varnothing})^+$ to $\mathcal{D}^*$.

**Theorem 2** *Let $\underline{P}$ be coherent on $\mathcal{X}|\mathcal{B}^{\varnothing}(\subset \mathcal{D}^*)$. Then, $\underline{E}(Z|A) = \underline{E}_{2c}^+(Z|A)$, $\forall Z|A \in \mathcal{D}^*$.*

*Proof* By Definitions 3 (d) and 4, $\underline{E}_{2c}^+(Z|A) = \sup L_{2c}^+(Z|A)$, where

$$\begin{aligned} L_{2c}^+(Z|A) = \{\alpha : &\sup\{B(X - \underline{P}(X|B)) - A(Z - \alpha)|A \vee B\} < 0, \\ &\text{for some } X|B \in (\mathcal{X}|\mathcal{B}^{\varnothing})^+\}. \end{aligned}$$

We show that $L_{2c}^+(Z|A) = L(Z|A)$.

In fact, if $\alpha \in L_{2c}^+(Z|A)$, then clearly $\alpha \in L^+(Z|A)$, hence $\alpha \in L(Z|A)$, because $L^+(Z|A) = L(Z|A)$ by Lemma 2.

Conversely, let $\alpha \in L(Z|A) = L_1(Z|A)$, by Proposition 2. Then, recalling (3), two distinct situations may occur:

(a) $\sup\{BX - A(Z - \alpha)|A \vee B\} < 0$, $X \in \mathcal{X}$, $B \in \mathcal{B}^{\varnothing}$, $\underline{P}(X|B) = 0$. Rewriting the supremum as $\sup\{B(X - \underline{P}(X|B)) - A(Z - \alpha)|A \vee B\} < 0$, then clearly $\alpha \in L_{2c}^+(Z|A)$.

(b) $\sup\{-A(Z - \alpha)|A\} < 0$. Since $0|A \in (\mathcal{X}|\mathcal{B}^{\varnothing})^+$, the supremum may be also written as $\sup\{A(0 - \underline{P}(0|A)) - A(Z - \alpha)|A\} < 0$, from which it is patent that $\alpha \in L_{2c}^+(Z|A)$.

Therefore, $L_{2c}^+(Z|A) = L(Z|A)$. The thesis follows taking the suprema. □

Theorem 2 assures that the natural extension and the 2-convex natural extension coincide, if $\underline{P}$ is coherent on $\mathcal{X}|\mathcal{B}^{\varnothing}$. Hence the 2-coherent natural extension $\underline{E}_2$ coincides with the former ones too, being sandwiched between them by Lemma 1.

Another result of the same kind is

**Theorem 3** *Let $\underline{P}$ be centered convex on $(\mathcal{X}|\mathcal{B}^{\varnothing})^+$. Then, $\underline{E}_c^+(Z|B) = \underline{E}_{2c}^+(Z|A)$, $\forall Z|A \in \mathcal{D}^*$.*

Finally, we can now establish the sandwich theorem:

**Theorem 4** (Sandwich Theorem) *Let $\underline{P}$ be coherent on $\mathcal{X}|\mathcal{B}^{\varnothing}$. Then $\underline{E}(Z|A) = \underline{E}_2(Z|A) = \underline{E}_c(Z|A) = \underline{E}_{2c}(Z|A)$, $\forall Z|A \in \mathcal{D}^*$.*

**Comment** The Sandwich Theorem ensures that the simpler 2-convex natural extension may be enough to compute the natural extension, or the convex natural extension, in the special case that the starting set is $\mathcal{X}|\mathcal{B}^{\varnothing}$. This seems to suggest that if $\underline{P}$ is initially assessed on a structured enough set and already coherent there, only the rather weak properties of (centered) 2-convexity really matter and need to be checked when looking for a least-committal coherent extension.

## 4 Conclusions

The results of the previous section show that the weak consistency notion of 2-convexity may be helpful in the important inferential problem of extending coherent or convex conditional (and unconditional) lower previsions. There remains to explore how this could be exploited in operational procedures, and whether the results can be applied to more general sets of conditional gambles than those in Definition 2. In our opinion, however, the present results already supply an additional motivation for further studying the interesting notion of 2-convexity.

**Acknowledgments** We acknowledge partial support by the FRA2015 grant 'Mathematical Models for Handling Risk and Uncertainty'.

## References

1. de Finetti B (1974) Theory of probability. Wiley
2. Pelessoni R, Vicig P (2005) Uncertainty modelling and conditioning with convex imprecise previsions. Int. J. Approx. Reason. 39(2–3):297–319
3. Pelessoni R, Vicig P (2009) Williams coherence and beyond. Int. J. Approx. Reason. 50(4):612–626
4. Pelessoni R, Vicig P (2015) Weak consistency for imprecise conditional previsions. In: Augustin T, Doria S, Miranda E, Quaeghebeur E (eds) Proceedings of the 9th international symposium on imprecise probability: theories and applications. Aracne Editrice
5. Pelessoni R, Vicig P (2016) 2-coherent and 2-convex conditional lower previsions. Submitted
6. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall
7. Williams PM (2007) Notes on conditional previsions. Int. J. Approx. Reason. 44(3):366–383

# Envelopes of Joint Probabilities with Given Marginals Under Absolute Continuity or Equivalence Constraints

**Davide Petturiti and Barbara Vantaggi**

**Abstract** The aim is to determine the envelopes of the class of joint probabilities (provided it is not empty) with assigned marginals, under the constraint of absolute continuity or equivalence with respect to a given reference measure.

## 1 Introduction

In many fields (such as economics, engineering and statistics) the pieces of information coming from different sources need to be aggregated in order to draw inferences. A distinguished problem is the so-called *marginal problem* in which, given the marginal distributions of some variables, one needs to establish whether there is a joint probability having the assigned marginal distributions [9, 12]. This kind of problems can occur, for instance, in the analysis of contingency tables, in mass transportation, in statistical matching and in misclassified variables problems. Recently, this problem [4] (see also [2]) has been faced by looking for the existence of a bivariate joint distribution having as marginals two given distributions and requiring some further condition: (a) the joint distribution is absolutely continuous with respect to a given measure, (b) the joint distribution is equivalent to a given measure. For the two problems above some remarkable existence results have been established in [4], under the logical independence assumption.

In this paper we consider the two aforementioned problems restricting to finite probability spaces, but allowing for logical relations. In general, a joint probability meeting condition (a) or (b), if it exists, is not unique. Here, we provide necessary and sufficient conditions for such classes not to be empty and, in this case, we give closed form expressions for their envelopes.

D. Petturiti (✉)
Dip. Matematica e Informatica, University of Perugia, Perugia, Italy
e-mail: davide.petturiti@dmi.unipg.it

B. Vantaggi
Dip. S.B.A.I.„ "La Sapienza" University of Rome, Rome, Italy
e-mail: barbara.vantaggi@sbai.uniroma1.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_49

Our results are interesting in order to draw inferences in multiple prior problems (see [6]) having marginal information, and distinguishing between structural zeroes (i.e., due to logical relations) and null probability events.

## 2 Finitely Additive Bivariate Marginal Problem

A set of events $\mathcal{G} = \{E_i\}_{i \in I}$ can always be embedded into a minimal Boolean algebra denoted as $\langle \mathcal{G} \rangle$ and said the *Boolean algebra generated* by $\mathcal{G}$.

A function $P : \mathcal{G} \to [0, 1]$ is a *coherent probability* [8] if and only if, for every $n \in \mathbb{N}$, every $E_{i_1} \ldots, E_{i_n} \in \mathcal{G}$ and every real numbers $s_1, \ldots, s_n$, the random gain (where $I_E$ denotes the indicator of an event $E$)

$$G = \sum_{j=1}^{n} s_j (I_{E_{i_j}} - P(E_{i_j})),$$

satisfies the following inequalities

$$\min_{C_r \in \mathcal{C}_{\mathcal{B}}} G(C_r) \leq 0 \leq \max_{C_r \in \mathcal{C}_{\mathcal{B}}} G(C_r),$$

where $\mathcal{C}_{\mathcal{B}} = \{C_1, \ldots, C_m\}$ is the set of atoms of $\mathcal{B} = \langle \{E_{i_1} \ldots, E_{i_n}\} \rangle$.

It is well-known that $P$ is coherent if and only if there exists a finitely additive probability $P'$ on $\langle \mathcal{G} \rangle$ such that $P'_{|\mathcal{G}} = P$. In general, the probability $P'$ is not unique but there is a class of finitely additive probability measures extending $P$.

In the following we consider two finitely additive probability measures $P_1$ on $\mathcal{A}_1$ and $P_2$ on $\mathcal{A}_2$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ are arbitrary Boolean algebras whose events are possibly linked by logical relations. The following theorem is a consequence of Theorem 3.6.1 in [5].

**Theorem 1** *The assessment* $\{P_1, P_2\}$ *on* $\mathcal{A}_1 \cup \mathcal{A}_2$ *is coherent if and only if it holds*

$$A_i \subseteq A_j \Longrightarrow P_i(A_i) \leq P_j(A_j), \quad \textit{for every } A_i \in \mathcal{A}_i, A_j \in \mathcal{A}_j, i \neq j. \tag{1}$$

In the case $\mathcal{A}_1$ and $\mathcal{A}_2$ are finite, condition (1) is equivalent to the one given in Theorem 1 of [11] (which holds for more than two marginal probability spaces), however, condition (1) is fairly easier to check.

Moreover, in the finite case, a probability $P$ on $\mathcal{B} = \langle \mathcal{A}_1 \cup \mathcal{A}_2 \rangle$ extending $\{P_1, P_2\}$ can be explicitly determined by solving the system with unknowns $x_{ij} = P(C_i \wedge D_j) \geq 0$ for every $C_i \wedge D_j \in \mathcal{C}_{\mathcal{B}}$

$$\mathcal{S} : \begin{cases} \displaystyle\sum_{\substack{C_i \wedge D_j \neq \emptyset \\ j=1,\ldots,m}} x_{ij} = P_1(C_i), \quad i = 1, \ldots, n, \\ \displaystyle\sum_{\substack{C_i \wedge D_j \neq \emptyset \\ i=1,\ldots,n}} x_{ij} = P_2(D_j), \quad j = 1, \ldots, m, \end{cases}$$

where $\mathcal{C}_{\mathcal{A}_1} = \{C_1, \dots, C_n\}$ and $\mathcal{C}_{\mathcal{A}_2} = \{D_1, \dots, D_m\}$, are the sets of atoms of $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, and $\mathcal{C}_{\mathcal{B}} = \{C_i \wedge D_j \neq \emptyset \,:\, i = 1, \dots, n,\, j = 1, \dots, m\}$ is the set of atoms of $\mathcal{B}$.

The resolution of system $\mathcal{S}$ can be faced through a recursive procedure, which progressively reduces the size of the two marginal assessments, as detailed below.

Up to a permutation, let $h \in \{1, \dots, n\}$ be such that $C_i \wedge \bigvee_{j=1}^{m-1} D_j \neq \emptyset$ for $i = 1, \dots, h$ and $C_i \wedge \bigvee_{j=1}^{m-1} D_j = \emptyset$ for $i = h+1, \dots, n$. Let $\tilde{\mathcal{A}}_1$ and $\tilde{\mathcal{A}}_2$ be two finite Boolean algebras with sets of atoms $\mathcal{C}_{\tilde{\mathcal{A}}_1} = \{\tilde{C}_1, \dots, \tilde{C}_h\}$ and $\mathcal{C}_{\tilde{\mathcal{A}}_2} = \{\tilde{D}_1, \dots, \tilde{D}_{m-1}\}$ such that $\tilde{C}_i \wedge \tilde{D}_j = \emptyset$ if and only if $C_i \wedge D_j = \emptyset$, for $i = 1, \dots, h, j = 1, \dots, m-1$. Consider the probability measures $\tilde{P}_1$ and $\tilde{P}_2$ on $\tilde{\mathcal{A}}_1$ and $\tilde{\mathcal{A}}_2$, respectively, whose distributions are $\tilde{P}_2(\tilde{D}_j) = \frac{P_2(D_j)}{\beta}$, for $j = 1, \dots, m-1$, $\tilde{P}_1(\tilde{C}_i) = \frac{P_1(C_i)-\alpha_i}{\beta}$, for $i = 1, \dots, h$, with $\beta = P_2(D_m^c)$ and $\bar{\alpha} = (\alpha_1, \dots, \alpha_n)$ a solution of the following system

$$\mathcal{S}' : \begin{cases} \sum_{i=1}^n \alpha_i = P_2(D_m), \\ h_i \leq \alpha_i \leq k_i \qquad \text{for } i = 1, \dots, n, \end{cases}$$

with $k_i = h_i = 0$ if $C_i \wedge D_m = \emptyset$, and $k_i = P_1(C_i) - \sum_{\substack{D_j \subseteq C_i \\ j=1,\dots,m-1}} P_2(D_j)$ and $h_i = \max\left\{0, P_1(C_i) - \sum_{\substack{D_j \wedge C_i \neq \emptyset \\ j=1,\dots,m-1}} P_2(D_j)\right\}$ otherwise. It is easily seen that condition (1) implies that $\mathcal{S}'$ has solution, moreover, for any such $\bar{\alpha}$ the global assessment $\{\tilde{P}_1, \tilde{P}_2\}$ on $\tilde{\mathcal{A}}_1 \cup \tilde{\mathcal{A}}_2$ still satisfies (1).

Thus the problem reduces to solve a smaller system $\mathcal{S}$ related to $\{\tilde{P}_1, \tilde{P}_2\}$. In particular, if the system $\mathcal{S}$ related to $\{\tilde{P}_1, \tilde{P}_2\}$ has a solution $(\tilde{x}_{ij})$, we get a solution of the system $\mathcal{S}$ related to $\{P_1, P_2\}$, setting $x_{ij} = \beta \tilde{x}_{ij}$ for $C_i \wedge D_j \neq \emptyset$, $i = 1, \dots, h$, $j = 1, \dots, m-1$, and $x_{im} = \alpha_i$ for $C_i \wedge D_m \neq \emptyset$, $i = 1, \dots, n$.

From now on we assume that $\{P_1, P_2\}$ is coherent, moreover, denote $\mathcal{B} = \langle \mathcal{A}_1 \cup \mathcal{A}_2 \rangle$ and consider

$$\mathcal{P} = \{P \,:\, P \text{ is a f.a. probability on } \mathcal{B} \text{ with } P_{|\mathcal{A}_i} = P_i,\, i = 1, 2\}.$$

The class $\mathcal{P}$ is a convex and compact subset of the space $[0, 1]^{\mathcal{B}}$ endowed with the product topology of pointwise convergence and the projection set on each element of $\mathcal{B}$ is a (possibly degenerate) closed interval. The pointwise envelopes

$$\underline{P} = \min \mathcal{P} \quad \text{and} \quad \overline{P} = \max \mathcal{P},$$

are known as *lower* and *upper probabilities*, respectively [13].

The following result provides a closed form expression for the envelopes $\underline{P}$ and $\overline{P}$ in terms of the Łukasiewicz t-norm $T_L$ and t-conorm $S_L$ defined (see [10]), for every $x, y \in [0, 1]$, as

$$T_L(x, y) = \max\{0, x + y - 1\} \quad \text{and} \quad S_L(x, y) = \min\{1, x + y\}.$$

**Theorem 2** *If $\{P_1, P_2\}$ on $\mathcal{A}_1 \cup \mathcal{A}_2$ is coherent, then the lower and upper envelopes $\underline{P}$ and $\overline{P}$ of the set $\mathcal{P}$ of coherent extensions of $\{P_1, P_2\}$ on $\mathcal{B}$ are such that, for every $B \in \mathcal{B}$*

$$\begin{aligned}\underline{P}(B) &= \max\left\{T_L(P_1(A_1), P_2(A_2)) \,:\, A_1 \wedge A_2 \subseteq B, A_i \in \mathcal{A}_i\right\},\\ \overline{P}(B) &= \min\left\{S_L(P_1(A_1), P_2(A_2)) \,:\, B \subseteq A_1 \vee A_2, A_i \in \mathcal{A}_i\right\}.\end{aligned}$$

*Proof* We prove the statement for $\overline{P}$ as that for $\underline{P}$ follows by duality. For every $B \in \mathcal{B}$, the fundamental theorem of probability [8] implies that

$$\overline{P}(B) = \min\left\{\overline{P}^{\mathcal{F}}(B) \,:\, \mathcal{F} \subseteq \mathcal{A}_1 \cup \mathcal{A}_2, \text{card}\, \mathcal{F} < \aleph_0\right\},$$

where $\overline{P}^{\mathcal{F}}(B)$ is the upper bound for the probability of $B$ obtained extending $P_{|\mathcal{F}}$ on $\mathcal{F} \cup \{B\}$. In particular, we can limit to finite subfamilies of the form $\mathcal{F} = \mathcal{A}'_1 \cup \mathcal{A}'_2$ with $\mathcal{A}'_i \subseteq \mathcal{A}_i$ finite subalgebra, for $i = 1, 2$. Coherence implies (see also Theorem 1 in [11]) that

$$\begin{aligned}\overline{P}^{\mathcal{F}}(B) &= \min\{P_1(A_1) + P_2(A_2) \,:\, B \subseteq A_1 \vee A_2, A_i \in \mathcal{A}'_i\}\\ &= \min\{1, \min\{P_1(A_1) + P_2(A_2) \,:\, B \subseteq A_1 \vee A_2, A_i \in \mathcal{A}'_i\}\}\\ &= \min\left\{S_L(P_1(A_1), P_2(A_2)) \,:\, B \subseteq A_1 \vee A_2, A_i \in \mathcal{A}'_i\right\},\end{aligned}$$

from which the thesis follows. □

If $\mathcal{B}$ is finite (and so also $\mathcal{A}_1$ and $\mathcal{A}_2$ are) with set of atoms $\mathcal{C}_{\mathcal{B}} = \{B_1, \ldots, B_s\}$, we provide a necessary and sufficient condition for the existence of a joint probability positive on $\mathcal{B}\setminus\{\emptyset\}$.

**Proposition 1** *If $\mathcal{A}_1$ and $\mathcal{A}_2$ are finite, then there exists $P \in \mathcal{P}$ such that $P(B) > 0$ for every $B \in \mathcal{B}\setminus\{\emptyset\}$ if and only if the following condition holds*

$$\min_{B_r \in \mathcal{C}_{\mathcal{B}}} \overline{P}(B_r) > 0. \tag{2}$$

*Proof* Condition (2) is trivially necessary so we prove its sufficiency. For every $B_r \in \mathcal{C}_{\mathcal{B}}$, there exists $P^r \in \mathcal{P}$ such that $P^r(B_r) = \overline{P}(B_r)$, thus by the finiteness of $\mathcal{B}$ the strict convex combination $P = \frac{1}{s}\sum_{r=1}^{s} P^r$ is plainly an element of $\mathcal{P}$ positive on $\mathcal{B}\setminus\{\emptyset\}$. □

## 3 Absolute Continuity and Equivalence Constraints

In this section, we assume that the two Boolean algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ are finite, moreover, besides the two probability measures $P_1$ and $P_2$ on $\mathcal{A}_1$ and $\mathcal{A}_2$, consider the reference measure $\mu : \mathcal{B} \to [0, +\infty)$, where $\mathcal{B} = \langle \mathcal{A}_1 \cup \mathcal{A}_2 \rangle$ having set of atoms $\mathcal{C}_{\mathcal{B}} = \{B_1, \ldots, B_s\}$.

Denote with $\mathcal{I} = \{B \in \mathcal{B} : \mu(B) = 0\}$ the ideal of $\mu$-null events in $\mathcal{B}$, which is the kernel of the canonical Boolean homomorphism $f : \mathcal{B} \to \mathcal{B}_{/\mathcal{I}}$, where, for every $B \in \mathcal{B}$, $\tilde{B} = f(B)$ denotes the equivalence class induced by $\mathcal{I}$. To avoid cumbersome notation, in what follows denote $\tilde{\mathcal{B}} = f(\mathcal{B})$ and $\tilde{\mathcal{A}}_i = f(\mathcal{A}_i)$, for $i = 1, 2$, for which it trivially follows $\tilde{\mathcal{B}} = \langle \tilde{\mathcal{A}}_1 \cup \tilde{\mathcal{A}}_2 \rangle$.

As usual, we say that a probability measure $P$ on $\mathcal{B}$ is *absolutely continuous* with respect to $\mu$, in symbol $P \ll \mu$, if and only if, for every $B \in \mathcal{B}$, $\mu(B) = 0$ implies $P(B) = 0$. Moreover, $P$ is *equivalent* to $\mu$, in symbol $P \sim \mu$, if and only if $P \ll \mu$ and $\mu \ll P$.

Assuming the coherence of $\{P_1, P_2\}$, consider the following subsets of the set $\mathcal{P}$ of probability measures on $\mathcal{B}$ with marginals $P_1$ and $P_2$:

- $\mathcal{P}^{\ll\mu} = \{P \in \mathcal{P} : P \ll \mu\}$;
- $\mathcal{P}^{\sim\mu} = \{P \in \mathcal{P} : P \sim \mu\}$.

As can be easily seen, the set $\mathcal{P}^{\ll\mu}$ is a closed subset of $[0, 1]^{\mathcal{B}}$, while $\mathcal{P}^{\sim\mu}$ is generally not [4].

The problems described above have been posed in [4], where $\mathcal{A}_1$ and $\mathcal{A}_2$ are assumed to be $\sigma$-fields of sets and $\mathcal{B} = \mathcal{A}_1 \otimes \mathcal{A}_2$, i.e., $\mathcal{A}_1$ and $\mathcal{A}_2$ are assumed to be logically independent. A related problem, where the probabilities in $\mathcal{P}$ are asked to be pointwise dominated by $\mu$ has been studied in [7].

In the following we give necessary and sufficient conditions for such sets not to be empty and, in this case, we provide a closed form expression for their envelopes.

**Theorem 3** *The following statements are equivalent:*

1. $\mathcal{P}^{\ll\mu} \neq \emptyset$;
2. $P_i$ *is constant on* $f^{-1}(\tilde{A}_i) \cap \mathcal{A}_i$, *for every* $\tilde{A}_i \in \tilde{\mathcal{A}}_i$, *and defining* $\tilde{P}_i(\tilde{A}_i) = P_i(A_i)$ *for* $A_i \in f^{-1}(\tilde{A}_i) \cap \mathcal{A}_i$, *for* $i = 1, 2$, $\{\tilde{P}_1, \tilde{P}_2\}$ *on* $\tilde{\mathcal{A}}_1 \cup \tilde{\mathcal{A}}_2$ *is coherent.*

*Proof 1.* $\Longrightarrow$ *2.* If $\mathcal{P}^{\ll\mu} \neq \emptyset$ and $P \in \mathcal{P}^{\ll\mu}$, then $P$ is constant on every $f^{-1}(\tilde{B})$, for every $\tilde{B} \in \tilde{\mathcal{B}}$, and, in particular, $P_i$ is constant on $f^{-1}(\tilde{A}_i) \cap \mathcal{A}_i$, for every $\tilde{A}_i \in \tilde{\mathcal{A}}_i$, for $i = 1, 2$. Defining $\tilde{P}(\tilde{B}) = P(B)$ for $B \in f^{-1}(\tilde{B})$ we get a probability on $\tilde{\mathcal{B}}$ which extends $\tilde{P}_1$ and $\tilde{P}_2$, and this implies the coherence of $\{\tilde{P}_1, \tilde{P}_2\}$.

*2.* $\Longrightarrow$ *1.* If $P_i$ is constant on $f^{-1}(\tilde{A}_i) \cap \mathcal{A}_i$, for every $\tilde{A}_i \in \tilde{\mathcal{A}}_i$, $\tilde{P}_i$ is defined as above, for $i = 1, 2$, and $\{\tilde{P}_1, \tilde{P}_2\}$ is coherent, then there exists a probability $\tilde{P}$ on $\tilde{\mathcal{B}}$ extending $\{\tilde{P}_1, \tilde{P}_2\}$. Defining, for every $B \in \mathcal{B}$, $P(B) = \tilde{P}(\tilde{B})$ we get a probability $P$ on $\mathcal{B}$ which extends $\{P_1, P_2\}$ and is such that, for every $B \in \mathcal{I}$, $P(B) = \tilde{P}(\tilde{B}) = \tilde{P}(\tilde{\emptyset}) = 0$, thus $P \in \mathcal{P}^{\ll\mu}$ and $\mathcal{P}^{\ll\mu} \neq \emptyset$. □

The following example shows an application of Theorem 3.

*Example 1* Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be the finite Boolean algebras with sets of atoms $\mathcal{C}_{\mathcal{A}_1} = \{C_1, C_2, C_3\}$ and $\mathcal{C}_{\mathcal{A}_2} = \{D_1, D_2, D_3\}$ with $C_1 \wedge D_1 = \emptyset$, and $P_1$ and $P_2$ the probability measures on $\mathcal{A}_1$ and $\mathcal{A}_2$ such that $P_1(C_1) = \frac{1}{2}$, $P_1(C_i) = \frac{1}{4}$, for $i = 2, 3$, and $P_2(D_j) = \frac{1}{3}$, for $j = 1, 2, 3$. Denote $\mathcal{B} = \langle \mathcal{A}_1 \cup \mathcal{A}_2 \rangle$, whose set of atoms is $\mathcal{C}_{\mathcal{B}} = \{B_{12}, B_{13}, B_{21}, B_{22}, B_{23}, B_{31}, B_{32}, B_{33}\}$, where $B_{ij} = C_i \wedge D_j$.

If we consider the reference measure $\mu$ on $\mathcal{B}$ whose distribution on $\mathcal{C}_{\mathcal{B}}$ is such that $\mu(B_{i3}) = 0$, for $i = 1, 2, 3$, and 1 otherwise, then $\mathcal{P}^{\ll\mu} = \emptyset$, since both $\emptyset$ and $D_3$ belong to $f^{-1}(\tilde{\emptyset}) \cap \mathcal{A}_2$, but $P_2(D_3) > P_2(\emptyset)$.

On the other hand, if we consider the reference measure $\mu'$ on $\mathcal{B}$ whose distribution on $\mathcal{C}_{\mathcal{B}}$ is such that $\mu'(B_{13}) = \mu'(B_{23}) = 0$ and 1 otherwise, then $P_i$ is constant on $f^{-1}(\tilde{A}_i) \cap \mathcal{A}_i$, for every $\tilde{A}_i \in \tilde{\mathcal{A}}_i$, but $\mathcal{P}^{\ll\mu'} = \emptyset$, since $\{\tilde{P}_1, \tilde{P}_2\}$ is not coherent as $\tilde{D}_3 \subseteq \tilde{C}_3$ and $\tilde{P}_2(\tilde{D}_3) > \tilde{P}_1(\tilde{C}_1)$. ■

**Theorem 4** *If $\mathcal{P}^{\ll\mu} \neq \emptyset$, and $\tilde{P}_i$ on $\tilde{\mathcal{A}}_i$, for $i = 1, 2$, is defined as in Theorem 3, then the lower and upper envelopes $\underline{P}^{\ll\mu} = \min \mathcal{P}^{\ll\mu}$ and $\overline{P}^{\ll\mu} = \max \mathcal{P}^{\ll\mu}$ are defined, for every $B \in \mathcal{B}$, as*

$$\underline{P}^{\ll\mu}(B) = \max\left\{T_L(\tilde{P}_1(\tilde{A}_1), \tilde{P}_2(\tilde{A}_2)) \,:\, \tilde{A}_1 \wedge \tilde{A}_2 \subseteq \tilde{B}, \tilde{A}_i \in \tilde{\mathcal{A}}_i\right\},$$
$$\overline{P}^{\ll\mu}(B) = \min\left\{S_L(\tilde{P}_1(\tilde{A}_1), \tilde{P}_2(\tilde{A}_2)) \,:\, \tilde{B} \subseteq \tilde{A}_1 \vee \tilde{A}_2, \tilde{A}_i \in \tilde{\mathcal{A}}_i\right\}.$$

*Proof* Let $\tilde{\mathcal{P}}$ be the set of probabilities on $\tilde{\mathcal{B}}$ extending $\{\tilde{P}_1, \tilde{P}_2\}$. By the proof of Theorem 3, every probability $P \in \mathcal{P}^{\ll\mu}$ is in bijection with a probability $\tilde{P} \in \tilde{\mathcal{P}}$ and it holds, for every $B \in \mathcal{B}$, $P(B) = \tilde{P}(\tilde{B})$. Hence, the conclusion follows by Theorem 2. □

In general, the fact $\mathcal{P}^{\ll\mu} \neq \emptyset$ does not imply $\mathcal{P}^{\sim\mu} \neq \emptyset$.

*Example 2* Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be the finite Boolean algebras with sets of atoms $\mathcal{C}_{\mathcal{A}_1} = \{C_1, C_2, C_3\}$ and $\mathcal{C}_{\mathcal{A}_2} = \{D_1, D_2, D_3\}$ such that $D_3 \wedge (C_1 \vee C_2) = C_3 \wedge (D_1 \vee D_2) = D_2 \wedge C_1 = \emptyset$ and $P_1$ and $P_2$ the probability measures on $\mathcal{A}_1$ and $\mathcal{A}_2$ such that $P_1(C_1) = P_2(D_1) = \frac{1}{2}$, $P_1(C_i) = P_2(D_i) = \frac{1}{4}$, for $i = 2, 3$. Consider $\mathcal{B} = \langle \mathcal{A}_1 \cup \mathcal{A}_2 \rangle$, whose set of atoms is $\mathcal{C}_{\mathcal{B}} = \{B_{11}, B_{21}, B_{22}, B_{33}\}$, where $B_{ij} = C_i \wedge D_j$. Let $\mu$ be the reference measure whose distribution on $\mathcal{C}_{\mathcal{B}}$ is constantly equal to 1. This implies that $\tilde{\mathcal{B}} = \mathcal{B}$, $\tilde{P}_i = P_i$, for $i = 1, 2$, and $\tilde{\mathcal{P}} = \mathcal{P}$.

The statement 2. of Theorem 3 applies since $\{P_1, P_2\}$ is coherent, so $\mathcal{P}^{\ll\mu} \neq \emptyset$ and it actually holds $\mathcal{P}^{\ll\mu} = \{P\}$ where $P$ is such that

| $\mathcal{C}_{\mathcal{B}}$ | $B_{11}$ | $B_{21}$ | $B_{22}$ | $B_{33}$ |
|---|---|---|---|---|
| $P$ | $\frac{1}{2}$ | 0 | $\frac{1}{4}$ | $\frac{1}{4}$ |

and this trivially implies $\mathcal{P}^{\sim\mu} = \emptyset$. ■

**Theorem 5** *The following statements are equivalent:*

1. *$\mathcal{P}^{\sim\mu} \neq \emptyset$;*
2. *$\mathcal{P}^{\ll\mu} \neq \emptyset$ and the upper envelope of the set $\tilde{\mathcal{P}}$ of probabilities on $\tilde{\mathcal{B}}$ extending $\{\tilde{P}_1, \tilde{P}_2\}$ satisfies (2).*

*Proof 1.* $\Longrightarrow$ *2.* If $\mathcal{P}^{\sim\mu} \neq \emptyset$ then there is a probability $P \in \mathcal{P}^{\ll\mu}$ such that, for every $B \in \mathcal{B}$, $P(B) = 0$ if and only if $B \in \mathcal{I}$. The probability $P$ is in bijection with a probability $\tilde{P}$ on $\tilde{\mathcal{B}}$ which extends $\{\tilde{P}_1, \tilde{P}_2\}$ and is positive on $\tilde{\mathcal{B}} \setminus \{\tilde{\emptyset}\}$, thus condition (2) holds.

*2.* $\Longrightarrow$ *1.* If $\mathcal{P}^{\ll\mu} \neq \emptyset$ and the upper envelope of the set $\tilde{\mathcal{P}}$ satisfies (2), then there is a probability $\tilde{P} \in \tilde{\mathcal{P}}$ which is positive on $\tilde{\mathcal{B}} \setminus \{\tilde{\emptyset}\}$. The probability $\tilde{P}$ is in bijection with a probability $P$ on $\mathcal{B}$ such that, for every $B \in \mathcal{B}$, $P(B) = 0$ if and only if $B \in \mathcal{I}$, thus $P \in \mathcal{P}^{\sim\mu}$. □

Theorems 3 and 5 are related to Theorems 7 and 13 given in [4] and to results in [1] and [3].

**Theorem 6** *If* $\mathcal{P}^{\sim\mu} \neq \emptyset$*, then it holds*

$$\underline{P}^{\sim\mu} = \inf \mathcal{P}^{\sim\mu} = \underline{P}^{\ll\mu} \text{ and } \overline{P}^{\sim\mu} = \sup \mathcal{P}^{\sim\mu} = \overline{P}^{\ll\mu}.$$

*Proof* We need to prove the statement for $\underline{P}^{\sim\mu}$ since the one for $\overline{P}^{\sim\mu}$ follows by duality. By Theorem 5, it holds $\mathcal{P}^{\sim\mu} \neq \emptyset$ if and only if $\mathcal{P}^{\ll\mu} \neq \emptyset$ and the upper envelope of $\tilde{\mathcal{P}}$ satisfies (2). Moreover, every probability in $\mathcal{P}^{\ll\mu}$ is in bijection with a probability in $\tilde{\mathcal{P}}$ and, in particular, every probability in $\mathcal{P}^{\sim\mu}$ is in bijection with a probability in $\tilde{\mathcal{P}}$ which is positive on $\tilde{\mathcal{B}} \setminus \{\tilde{\emptyset}\}$. The set $\tilde{\mathcal{P}}$ is a convex compact subset of $[0, 1]^{\tilde{\mathcal{B}}}$ endowed with the product topology of pointwise convergence such that, $\mathrm{proj}_{\tilde{B}}(\tilde{\mathcal{P}}) = [\underline{\pi}_{\tilde{B}}, \overline{\pi}_{\tilde{B}}] \subseteq [0, 1]$, for every $\tilde{B} \in \tilde{\mathcal{B}}$.

We show that the pointwise infimum of the set of positive probabilities in $\tilde{\mathcal{P}}$ coincides with the pointwise infimum of the whole $\tilde{\mathcal{P}}$. Let $\tilde{P} \in \tilde{\mathcal{P}}$ be positive on $\tilde{\mathcal{B}} \setminus \{\tilde{\emptyset}\}$ and fix $\tilde{B} \in \tilde{\mathcal{B}}$ for which we have $\underline{\pi}_{\tilde{B}} \leq \tilde{P}(\tilde{B})$. If $\underline{\pi}_{\tilde{B}} = \tilde{P}(\tilde{B})$ then the statement trivially holds, otherwise, suppose $\underline{\pi}_{\tilde{B}} < \tilde{P}(\tilde{B})$ and let $\tilde{Q} \in \tilde{\mathcal{P}}$ be such that $\tilde{Q}(\tilde{B}) = \underline{\pi}_{\tilde{B}}$. For every $\epsilon \in (\underline{\pi}_{\tilde{B}}, \tilde{P}(\tilde{B}))$, the strict convex combination $\tilde{P}^{\epsilon} = \alpha\tilde{Q} + (1 - \alpha)\tilde{P}$ with $\alpha = \frac{\tilde{P}(\tilde{B}) - \epsilon}{1 - \underline{\pi}_{\tilde{B}}}$ is a probability in $\tilde{\mathcal{P}}$ which is positive on $\tilde{\mathcal{B}} \setminus \{\tilde{\emptyset}\}$ and such that $\tilde{P}^{\epsilon}(\tilde{B}) = \epsilon$, and this concludes the proof. □

Since the class $\mathcal{P}^{\sim\mu}$ is not necessarily closed, its envelopes could not be attained pointwise by any of its elements, as shown in the following example.

*Example 3* Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be the logically independent finite Boolean algebras with sets of atoms $\mathcal{C}_{\mathcal{A}_1} = \{C_1, C_2, C_3\}$ and $\mathcal{C}_{\mathcal{A}_2} = \{D_1, D_2, D_3\}$, and $P_1$ and $P_2$ the probability measures on $\mathcal{A}_1$ and $\mathcal{A}_2$ such that $P_1(C_1) = P_1(C_2) = P_1(C_3) = \frac{1}{3}$, $P_2(D_1) = P_2(D_3) = \frac{1}{4}$ and $P_2(D_2) = \frac{1}{2}$. Consider $\mathcal{B} = \langle \mathcal{A}_1 \cup \mathcal{A}_2 \rangle$ which has set of atoms $\mathcal{C}_{\mathcal{B}} = \{B_{ij} = C_i \wedge D_j \,:\, i, j = 1, 2, 3\}$, and let $\mu$ be the reference measure on $\mathcal{B}$ whose distribution on $\mathcal{C}_{\mathcal{B}}$ is such that $\mu(B_{ii}) = 0$ for $i = 1, 2, 3$ and 1 otherwise.

The algebras $\tilde{\mathcal{A}}_1$ and $\tilde{\mathcal{A}}_2$ have sets of atoms $\mathcal{C}_{\tilde{\mathcal{A}}_1} = \{\tilde{C}_1, \tilde{C}_2, \tilde{C}_3\}$ and $\mathcal{C}_{\tilde{\mathcal{A}}_2} = \{\tilde{D}_1, \tilde{D}_2, \tilde{D}_3\}$ such that $\tilde{C}_i \wedge \tilde{D}_i = \tilde{\emptyset}$ for $i = 1, 2, 3$. The statement *2.* of Theorem 3 applies since $\{\tilde{P}_1, \tilde{P}_2\}$ is coherent, so $\mathcal{P}^{\ll\mu} \neq \emptyset$ and it actually holds $\mathcal{P}^{\ll\mu} = \left\{P^{\gamma} \,:\, \gamma \in \left[0, \frac{1}{6}\right]\right\}$ where each $P^{\gamma} \in \mathcal{P}^{\ll\mu}$ is such that

| $\mathcal{C}_{\mathcal{B}}$ | $B_{11}$ | $B_{12}$ | $B_{13}$ | $B_{21}$ | $B_{22}$ | $B_{23}$ | $B_{31}$ | $B_{32}$ | $B_{33}$ |
|---|---|---|---|---|---|---|---|---|---|
| $P^{\gamma}$ | 0 | $\frac{1}{3} - \gamma$ | $\gamma$ | $\frac{1}{12} + \gamma$ | 0 | $\frac{1}{4} - \gamma$ | $\frac{1}{6} - \gamma$ | $\frac{1}{6} + \gamma$ | 0 |

Simple computations show that the upper envelope of the set $\tilde{\mathcal{P}}$ of probabilities extending $\{\tilde{P}_1, \tilde{P}_2\}$ on $\tilde{\mathcal{B}}$ satisfies (2), so by Theorem 5 we have that $\mathcal{P}^{\sim\mu} \neq \emptyset$ and it actually holds $\mathcal{P}^{\sim\mu} = \left\{P^{\gamma} \in \mathcal{P}^{\ll\mu} \,:\, \gamma \in \left(0, \frac{1}{6}\right)\right\}$. Finally, it is trivially seen that $\underline{P}^{\ll\mu} = \underline{P}^{\sim\mu}$ and $\overline{P}^{\ll\mu} = \overline{P}^{\sim\mu}$. ■

**Acknowledgments** Work partially supported by INdAM-GNAMPA through the Project 2015 U2015/000418 and by the Italian MIUR PRIN 2010-11 2010FP79LR_003.

## References

1. Berti P, Pratelli L, Rigo P (2010) Finitely additive equivalent martingale measures. J Theor Prob 26(1):46–57
2. Berti P, Pratelli L, Rigo P (2014) A unifying view on some problems in probability and statistics. Stat Methods Appl 23(4):483–500
3. Berti P, Pratelli L, Rigo P (2015) Two versions of the fundamental theorem of asset pricing. Electron J Prob 20(34):1–21
4. Berti P, Pratelli L, Rigo P, Spizzichino F (2015) Equivalent or absolutely continuous probability measures with given marginals. Depend Model 3:47–58
5. Bhaskara Rao K, Bhaskara Rao M (1983) Theory of Charges: A Study of Finitely Additive Measures. Academic Press
6. Coletti G, Petturiti D, Vantaggi B (2014) Bayesian inference: the role of coherence to deal with a prior belief function. Stat Methods Appl 23(4):519–545
7. Dall'Aglio G (1972) Fréchet classes and compatibility of distribution functions, Symp. Math., vol 9, Academic Press London, New York, pp 131–150
8. de Finetti B (1974) Theory of probability, vol 1–2. John Wiley & Sons
9. Fréchet M (1951) Sur les tableaux de corrélation dont les marges sont données. Ann Univ Lyon Sci 4:13–84
10. Klement E, Mesiar R, Pap E (2000) Triangular Norms, Trends in Logic, vol 8. Kluwer Academic Publishers
11. Rüschendorf L (1991) Fréchet-Bounds and Their Applications, vol 67. Mathematics and Its Applications. Springer, Netherlands
12. Strassen V (1965) The existence of probability measures with given marginals. Ann Math Stat 36(2):423–439
13. Williams P (2007) Note on conditional previsions. Int J Approx Reason 44:366–383

# Square of Opposition Under Coherence

**Niki Pfeifer and Giuseppe Sanfilippo**

**Abstract** Various semantics for studying the square of opposition have been proposed recently. So far, only [14] studied a probabilistic version of the square where the sentences were interpreted by (negated) defaults. We extend this work by interpreting *sentences* by imprecise (set-valued) probability assessments on a sequence of conditional events. We introduce the *acceptability* of a sentence within coherence-based probability theory. We analyze the relations of the square in terms of acceptability and show how to construct probabilistic versions of the square of opposition by forming suitable tripartitions. Finally, as an application, we present a new square involving generalized quantifiers.

## 1 Introduction

There is a long history of investigations on the square of opposition spanning over two millenia [1, 19]. A *square of opposition* (SOP) represents logical relations among basic sentence types in a diagrammatic way. The basic sentence types, traditionally denoted by $A$ (universal affirmative: “Every $S$ is $P$”), $E$ (universal negative: “No $S$ is $P$”), $I$ (particular affirmative: “Some $S$ are $P$”), and $O$ (particular negative: “Some $S$ are not $P$”), constitute the corners of the square, and the logical relations—*contradiction*, *contrarity*, *subalternation*, and *sub-contrarity*—form the diagonals and the sides of the square. Recently, the square has been investigated from various semantic points of view (see, e.g., [1, 9]). The present paper deepens the probabilistic analysis of the SOP under coherence given in [14]. After preliminary notions (Sect. 2), we introduce, based on g-coherence, a (probabilistic) notion of sentences and their

Shared first authorship (both authors contributed equally to this work).

N. Pfeifer
Ludwig-Maximilians-University Munich, Munich, Germany
e-mail: niki.pfeifer@lmu.de

G. Sanfilippo (✉)
University of Palermo, Palermo, Italy
e-mail: giuseppe.sanfilippo@unipa.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_50

acceptability and show how to construct squares of opposition under coherence from suitable tripartitions (Sect. 3). Then, we present an application of our square to the study of generalized quantifiers (Sect. 4). Section 5 concludes the paper by some remarks on future work.

## 2 Preliminary Notions

Given two events $E$ and $H$, with $H \neq \bot$, the *conditional event* $E|H$ is defined as a three-valued logical entity which is *true* if $EH$ (i.e., $E \wedge H$) is true, *false* if $\overline{E}H$ is true, and *void* if $H$ is false. Given a finite sequence of $n \geq 1$ conditional events $\mathcal{F} = (E_1|H_1, \dots, E_n|H_n)$, we denote by $\mathcal{P}$ *any precise* probability assessment $\mathcal{P} = (p_1, \dots, p_n)$ on $\mathcal{F}$, where $p_j = p(E_j|H_j) \in [0, 1]$, $j = 1, \dots, n$. Moreover, we denote by $\Pi$ the set of *all coherent precise* assessments on $\mathcal{F}$. The coherence-based approach to probability and to other uncertain measures has been adopted by many authors (see, e.g., [2–4, 6–8, 10, 12, 16, 17, 21–23]); we therefore recall only selected key features of coherence in this paper. We recall that when there are no logical relations among the events $E_1, H_1, \dots, E_n, H_n$ involved in $\mathcal{F}$, that is $E_1, H_1, \dots, E_n, H_n$ are logically independent, then the set $\Pi$ associated with $\mathcal{F}$ is the whole unit hypercube $[0, 1]^n$. If there are logical relations, then the set $\Pi$ *could be* a strict subset of $[0, 1]^n$. As it is well known $\Pi \neq \emptyset$; therefore, $\emptyset \neq \Pi \subseteq [0, 1]^n$. If not stated otherwise, we do not make any assumptions concerning logical independence.

**Definition 1** An *imprecise, or set-valued, assessment* $\mathcal{I}$ on a family of conditional events $\mathcal{F}$ is a (possibly empty) set of precise assessments $\mathcal{P}$ on $\mathcal{F}$.

Definition 1 states that an *imprecise (probability) assessment* $\mathcal{I}$ on a sequence of $n$ conditional events $\mathcal{F}$ is just a (possibly empty) subset of $[0, 1]^n$ ([11, 13, 14]). For instance, think about an agent (like Pythagoras) who considers only rational numbers to evaluate the probability of an event $E|H$. Pythagoras' evaluation can be represented by the imprecise assessment $\mathcal{I} = [0, 1] \cap \mathbb{Q}$ on $E|H$. Moreover, a constraint like $p(E|H) > 0$ can be represented by $\mathcal{I} = ]0, 1]$.

Given an imprecise assessment $\mathcal{I}$ we denote by $\overline{\mathcal{I}}$ the *complementary imprecise assessment* of $\mathcal{I}$, i.e. $\overline{\mathcal{I}} = [0, 1]^n \setminus \mathcal{I}$. We now recall the notions of g-coherence and total coherence in the general case of imprecise (in the sense of set-valued) probability assessments [14].

**Definition 2** (*g-coherence*) Given a sequence of $n$ conditional events $\mathcal{F}$. An imprecise assessment $\mathcal{I} \subseteq [0, 1]^n$ on $\mathcal{F}$ is *g-coherent* iff there exists a coherent precise assessment $\mathcal{P}$ on $\mathcal{F}$ such that $\mathcal{P} \in \mathcal{I}$.

**Definition 3** (*t-coherence*) An imprecise assessment $\mathcal{I}$ on $\mathcal{F}$ is *totally coherent* (t-coherent) iff the following two conditions are satisfied: (i) $\mathcal{I}$ is non-empty; (ii) if $\mathcal{P} \in \mathcal{I}$, then $\mathcal{P}$ is a coherent precise assessment on $\mathcal{F}$.

**Definition 4** (*t-coherent part*) Given a sequence of $n$ conditional events $\mathcal{F}$. Let $\Pi$ be the set of all coherent assessments on $\mathcal{F}$. We denote by $\pi : \wp([0,1]^n) \rightarrow \wp(\Pi)$ the function defined by $\pi(\mathcal{I}) = \Pi \cap \mathcal{I}$, for any imprecise assessment $\mathcal{I} \in \wp([0,1]^n)$. Moreover, for each subset $\mathcal{I} \in \wp([0,1]^n)$ we call $\pi(\mathcal{I})$ the *t-coherent part* of $\mathcal{I}$.

Of course, if $\pi(\mathcal{I}) \neq \emptyset$, then $\mathcal{I}$ is g-coherent and $\pi(\mathcal{I})$ is t-coherent.

## 3 From Imprecise Assessments to the Square of Opposition

In this section we consider imprecise assessments on a given sequence $\mathcal{F}$ of $n$ conditional events. In our approach, a sentence $s$ is a pair $(\mathcal{F}, \mathcal{I})$, where $\mathcal{I} \subseteq [0,1]^n$ is an imprecise assessment on $\mathcal{F}$. We introduce the following equivalence relation under t-coherence:

**Definition 5** Given two sentences $s_1 : (\mathcal{F}, \mathcal{I}_1)$ and $s_2 : (\mathcal{F}, \mathcal{I}_2)$, $s_1$ and $s_2$ are *equivalent (under t-coherence)*, denoted by $s_1 \equiv s_2$, iff $\pi(\mathcal{I}_1) = \pi(\mathcal{I}_2)$.

**Definition 6** Given three sentences $s : (\mathcal{F}, \mathcal{I})$, $s_1 : (\mathcal{F}, \mathcal{I}_1)$, and $s_2 : (\mathcal{F}, \mathcal{I}_2)$. We define: $s_1 \wedge s_2 : (\mathcal{F}, \mathcal{I}_1 \cap \mathcal{I}_2)$ (conjunction); $s_1 \vee s_2 : (\mathcal{F}, \mathcal{I}_1 \cup \mathcal{I}_2)$ (disjunction); $\bar{s} : (\mathcal{F}, \overline{\mathcal{I}})$, where $\overline{\mathcal{I}} = [0,1]^n \setminus \mathcal{I}$ (negation).

*Remark 1* As the basic operations among sentences are defined by set-theoretical operations, they inherit the corresponding properties (including associativity, commutativity, De Morgan Laws, etc.). Moreover, as $\pi(\mathcal{I}_1 \cap \mathcal{I}_2) = \pi(\mathcal{I}_1) \cap \pi(\mathcal{I}_2)$, by setting $s_1^* = (\mathcal{F}, \pi(\mathcal{I}_1)), s_2^* = (\mathcal{F}, \pi(\mathcal{I}_2))$ and $(s_1 \wedge s_2)^* : (\mathcal{F}, \pi(\mathcal{I}_1 \cap \mathcal{I}_2))$, it follows that $(s_1 \wedge s_2) \equiv (s_1 \wedge s_2)^* \equiv s_1^* \wedge s_2^*$. Likewise, $s_1 \vee s_2 \equiv (s_1 \vee s_2)^* \equiv s_1^* \vee s_2^*$.

As we interpret the basic sentence types involved in the SOP by imprecise probability assessments on sequences of conditional events, we will introduce the following notion of acceptability, which serves as a semantic bridge between basic sentence types and imprecise assessments:

**Definition 7** A sentence $s : (\mathcal{F}, \mathcal{I})$ is (resp., is not) *acceptable* iff the assessment $\mathcal{I}$ on $\mathcal{F}$ is (resp., is not) g-coherent, i.e. $\pi(\mathcal{I})$ is not (resp., is) empty.

*Remark 2* If $s_1 \wedge s_2$ is acceptable, then $s_1$ is acceptable and $s_2$ is acceptable. However, the converse does not hold, indeed $s_1 : (E|H, \{1\})$ is acceptable and $s_2 : (E|H), \{0\})$ is acceptable, but $s_1 \wedge s_2 : (E|H, \emptyset)$ is not acceptable (as $\pi(\emptyset) = \emptyset$).

**Definition 8** Given two sentences $s_1 : (\mathcal{F}, \mathcal{I}_1)$ and $s_2 : (\mathcal{F}, \mathcal{I}_2)$, we say, under coherence: $s_1$ and $s_2$ are *contraries* iff the sentence $s_1 \wedge s_2$ is not acceptable[1]; $s_1$ and $s_2$ are *subcontraries* iff $\bar{s}_1 \wedge \bar{s}_2$ is not acceptable; $s_1$ and $s_2$ are *contradictories* iff $s_1$ and $s_2$ are both, contraries and subcontraries; $s_2$ is a *subaltern* of $s_1$ iff the sentence $s_1 \wedge \bar{s}_2$ is not acceptable.

[1] Some definitions of contrariety additionally require that "$s_1$ and $s_2$ can both be acceptable." For reasons stated in [14], we omit this additional requirement. Similarly, *mutatis mutandis*, in our definition of subcontrariety.

*Remark 3* We observe that $s_1 \wedge \bar{s}_2$ is not acceptable iff $\Pi \cap (\mathcal{I}_1 \cap \overline{\mathcal{I}}_2) = \emptyset$, which also amounts to say that $\Pi \cap \mathcal{I}_1 \subseteq \Pi \cap \mathcal{I}_2$. Moreover, if $s_1$ is not acceptable, that is $\Pi \cap \mathcal{I}_1 = \emptyset$, then any sentence $s_2$ is a subaltern of $s_1$. For instance, the sentence $s_1 : (E|\overline{E}, 1)$ is not acceptable and then any sentence $s_2 : (E|\overline{E}, \mathcal{I})$, where $\mathcal{I} \subseteq [0, 1]$, is a subaltern of $s_1$.

**Definition 9** Let $s_k : (\mathcal{F}, \mathcal{I}_k)$, $k = 1, 2, 3, 4$, be four sentences. We call the ordered quadruple $(s_1, s_2, s_3, s_4)$ a *square of opposition* (under coherence), iff the following relations among the four sentences hold:

(a) $s_1$ and $s_2$ are contraries, i.e., $\pi(\mathcal{I}_1) \cap \pi(\mathcal{I}_2) = \emptyset$;
(b) $s_3$ and $s_4$ are subcontraries, i.e., $\pi(\mathcal{I}_3) \cup \pi(\mathcal{I}_4) = \Pi$;
(c) $s_1$ and $s_4$ are contradictories, i.e., $\pi(\mathcal{I}_1) \cap \pi(\mathcal{I}_4) = \emptyset$ and $\pi(\mathcal{I}_1) \cup \pi(\mathcal{I}_4) = \Pi$;
$s_2$ and $s_3$ are contradictories, i.e., $\pi(\mathcal{I}_2) \cap \pi(\mathcal{I}_3) = \emptyset$ and $\pi(\mathcal{I}_2) \cup \pi(\mathcal{I}_3) = \Pi$;
(d) $s_3$ is a subaltern of $s_1$, i.e., $\pi(\mathcal{I}_1) \subseteq \pi(\mathcal{I}_3)$;
$s_4$ is a subaltern of $s_2$, i.e., $\pi(\mathcal{I}_2) \subseteq \pi(\mathcal{I}_4)$.

*Remark 4* Based on Definition 9, we observe that in order to verify if a quadruple of sentences $(s_1, s_2, s_3, s_4)$, where $s_k : (\mathcal{F}, \mathcal{I}_k)$ and $k = 1, 2, 3, 4$, is a SOP, it is necessary and sufficient to check that the quadruple $(s'_1, s'_2, s'_3, s'_4)$, where $s'_k = (\mathcal{F}, \mathcal{I}'_k)$, $\mathcal{I}'_k = \pi(\mathcal{I}_k)$, is a SOP. Then, we say that two squares $(s_1, s_2, s_3, s_4)$ and $(s'_1, s'_2, s'_3, s'_4)$ *coincide* iff $\pi(\mathcal{I}_k) = \pi(\mathcal{I}'_k)$ for each $k$. Moreover, based on Definition 9, we observe that $(s_1, s_2, s_3, s_4)$ is a SOP iff $(s_2, s_1, s_4, s_3)$ is a SOP.

**Definition 10** An (ordered) *tripartition* of a set $\mathfrak{S}$ is a triple $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$, where $\mathcal{D}_1, \mathcal{D}_2$, and $\mathcal{D}_3$ are subsets of $\mathfrak{S}$, such that the following conditions are satisfied: (i) $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$, $i \neq j$ for all $i, j = 1, 2, 3$; (ii); $\mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3 = \mathfrak{S}$.

**Theorem 1** *Given any sequence of n conditional events $\mathcal{F}$ and a quadruple $(s_1, s_2, s_3, s_4)$ of sentences, with $s_k : (\mathcal{F}, \mathcal{I}_k)$, $k = 1, 2, 3, 4$. Define $\mathcal{D}_1 = \pi(\mathcal{I}_1)$, $\mathcal{D}_2 = \pi(\mathcal{I}_2)$, and $\mathcal{D}_3 = \pi(\mathcal{I}_3) \cap \pi(\mathcal{I}_4)$. Then, the quadruple $(s_1, s_2, s_3, s_4)$ is a SOP if and only if $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$ is a tripartition of (the non-empty set) $\Pi$ such that: $\pi(\mathcal{I}_3) = \mathcal{D}_1 \cup \mathcal{D}_3$, $\pi(\mathcal{I}_4) = \mathcal{D}_2 \cup \mathcal{D}_3$.*

*Proof* $(\Rightarrow)$. We assume that $\mathcal{D}_1 = \pi(\mathcal{I}_1)$, $\mathcal{D}_2 = \pi(\mathcal{I}_2)$, and $\mathcal{D}_3 = \pi(\mathcal{I}_3) \cap \pi(\mathcal{I}_4)$. Of course, $\mathcal{D}_i \subseteq \Pi$, $i = 1, 2, 3$. We now prove that: $(i)$ $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$; $(ii)$ $\mathcal{D}_3 = \Pi \setminus (\mathcal{D}_1 \cup \mathcal{D}_2)$. $(i)$ From condition (a) in Definition 9, as $s_1$ and $s_2$ are contraries, it follows that $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$. $(ii)$ We first prove that $\mathcal{D}_3 \subseteq \Pi \setminus (\mathcal{D}_1 \cup \mathcal{D}_2)$. This trivially follows when $\mathcal{D}_3 = \emptyset$. If $\mathcal{D}_3 \neq \emptyset$, then let $x \in \mathcal{D}_3 = \pi(\mathcal{I}_3) \cap \pi(\mathcal{I}_4)$. As $x \in \pi(\mathcal{I}_3)$, from condition (c) in Definition 9, we obtain $x \notin \pi(\mathcal{I}_2)$. Likewise, as $x \in \pi(\mathcal{I}_4)$, from condition (c) in Definition 9, we obtain $x \notin \pi(\mathcal{I}_1)$. Then, $x \in \Pi$ and $x \notin (\pi(\mathcal{I}_1) \cup \pi(\mathcal{I}_2))$, that is $x \in \Pi \setminus (\mathcal{D}_1 \cup \mathcal{D}_2)$. We now prove that $\Pi \setminus (\mathcal{D}_1 \cup \mathcal{D}_2) \subseteq \mathcal{D}_3$. This trivially follows when $\Pi \setminus (\mathcal{D}_1 \cup \mathcal{D}_2) = \emptyset$. If $\Pi \setminus (\mathcal{D}_1 \cup \mathcal{D}_2) \neq \emptyset$, let $x \in \Pi \setminus (\pi(\mathcal{I}_1) \cup \pi(\mathcal{I}_2))$. As $x \in \Pi \setminus \pi(\mathcal{I}_1)$, from condition (c) in Definition 9, we obtain $x \in \pi(\mathcal{I}_4)$. Likewise, as $x \in \Pi \setminus \pi(\mathcal{I}_2)$ from condition (c) in Definition 9, we obtain $x \in \pi(\mathcal{I}_3)$. Then, $x \in (\pi(\mathcal{I}_3) \cap \pi(\mathcal{I}_4)) = \mathcal{D}_3$. Therefore $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$ is a tripartition of $\Pi$. By our assumption, $\pi(\mathcal{I}_1) = \mathcal{D}_1$ and $\pi(\mathcal{I}_2) = \mathcal{D}_2$. We observe that

$\pi(\mathcal{I}_3) \cap \mathcal{D}_3 = \mathcal{D}_3$; moreover, from conditions (c) and (d), we obtain $\pi(\mathcal{I}_3) \cap \mathcal{D}_2 = \pi(\mathcal{I}_3) \cap \pi(\mathcal{I}_2) = \emptyset$ and $\pi(\mathcal{I}_3) \cap \mathcal{D}_1 = \pi(\mathcal{I}_1) \cap \pi(\mathcal{I}_3) = \pi(\mathcal{I}_1) = \mathcal{D}_1$; then $\pi(\mathcal{I}_3) = \pi(\mathcal{I}_3) \cap (\mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3) = \mathcal{D}_1 \cup \mathcal{D}_3$. Likewise, we observe that $\pi(\mathcal{I}_4) \cap \mathcal{D}_3 = \mathcal{D}_3$; moreover, from conditions (c),(d) in Definition 9, we obtain $\mathcal{D}_1 \cap \pi(\mathcal{I}_4) = \pi(\mathcal{I}_1) \cap \pi(\mathcal{I}_4) = \emptyset$ and $\mathcal{D}_2 \cap \pi(\mathcal{I}_4) = \pi(\mathcal{I}_2) \cap \pi(\mathcal{I}_4) = \pi(\mathcal{I}_2) = \mathcal{D}_2$; then $\pi(\mathcal{I}_4) = \pi(\mathcal{I}_4) \cap (\mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3) = \mathcal{D}_2 \cup \mathcal{D}_3$.

($\Leftarrow$) Assume that $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$, where $\mathcal{D}_1 = \pi(\mathcal{I}_1)$, $\mathcal{D}_2 = \pi(\mathcal{I}_2)$, $\mathcal{D}_3 = \pi(\mathcal{I}_3) \cap \pi(\mathcal{I}_4)$, is a tripartition of $\Pi$ such that $\mathcal{D}_1 \cup \mathcal{D}_3 = \pi(\mathcal{I}_3)$ and $\mathcal{D}_2 \cup \mathcal{D}_3 = \pi(\mathcal{I}_4)$, we prove that the quadruple $(s_1, s_2, s_3, s_4)$ satisfies conditions (a), (b), (c), and (d) in Definition 9. We observe that $\pi(\mathcal{I}_1) \cap \pi(\mathcal{I}_2) = \mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$, which coincides with (a). Condition (b) is satisfied because $\pi(\mathcal{I}_3) \cup \pi(\mathcal{I}_4) = \mathcal{D}_1 \cup \mathcal{D}_3 \cup \mathcal{D}_2 \cup \mathcal{D}_3 = \Pi$. Moreover, $\pi(\mathcal{I}_1) \cap \pi(\mathcal{I}_4) = \mathcal{D}_1 \cap (\mathcal{D}_2 \cup \mathcal{D}_3) = \emptyset$ and $\pi(\mathcal{I}_1) \cup \pi(\mathcal{I}_4) = \mathcal{D}_1 \cup (\mathcal{D}_2 \cup \mathcal{D}_3) = \Pi$; likewise, $\pi(\mathcal{I}_2) \cap \pi(\mathcal{I}_3) = \mathcal{D}_2 \cap (\mathcal{D}_1 \cup \mathcal{D}_3) = \emptyset$ and $\pi(\mathcal{I}_2) \cup \pi(\mathcal{I}_3) = \mathcal{D}_2 \cup (\mathcal{D}_1 \cup \mathcal{D}_3) = \Pi$. Thus, the conditions in (c) are satisfied. Finally, $\pi(\mathcal{I}_1) = \mathcal{D}_1 \subseteq \mathcal{D}_1 \cup \mathcal{D}_3 = \pi(\mathcal{I}_3)$ and $\pi(\mathcal{I}_2) = \mathcal{D}_2 \subseteq \mathcal{D}_2 \cup \mathcal{D}_3 = \pi(\mathcal{I}_4)$ which satisfy conditions in (d). □

A method to construct a SOP by starting from a tripartition of $\Pi$ is given in the following result (see also [9]).

**Corollary 1** *Given any sequence of $n$ conditional events $\mathcal{F}$ and a tripartition $(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$ of $\Pi$, then the quadruple $(s_1, s_2, s_3, s_4)$, with $s_k : (\mathcal{F}, \mathcal{I}_k)$, $k = 1, 2, 3, 4$ and $\pi(\mathcal{I}_1) = \mathcal{D}_1$, $\pi(\mathcal{I}_2) = \mathcal{D}_2$, $\pi(\mathcal{I}_3) = \mathcal{D}_1 \cup \mathcal{D}_3$, $\pi(\mathcal{I}_4) = \mathcal{D}_2 \cup \mathcal{D}_3$ is a SOP.*

*Proof* The proof immediately follows by observing $\pi(\mathcal{I}_3) \cap \pi(\mathcal{I}_4) = \mathcal{D}_3$ and by the ($\Leftarrow$) side proof of Theorem 1. □

The following result allows to construct a SOP by starting from a tripartition of the whole set $[0, 1]^n$:

**Corollary 2** *Given a tripartition $(\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3)$ of $[0, 1]^n$, let $\mathcal{I}_1 = \mathcal{B}_1$, $\mathcal{I}_2 = \mathcal{B}_2$, $\mathcal{I}_3 = \mathcal{B}_1 \cup \mathcal{B}_3$, and $\mathcal{I}_4 = \mathcal{B}_2 \cup \mathcal{B}_3$. For any sequence of $n$ conditional events $\mathcal{F}$, the quadruple $(s_1, s_2, s_3, s_4)$, where $s_k : (\mathcal{F}, \mathcal{I}_k)$, $k = 1, 2, 3, 4$, is a SOP.*

*Proof* Let $\mathcal{F}$ be any sequence of $n$ conditional events and $\Pi$ be the associated set of all coherent precise assessments. We set $\mathcal{D}_i = \pi(\mathcal{B}_i)$, $i = 1, 2, 3$. Of course, $(\pi(\mathcal{B}_1), \pi(\mathcal{B}_2), \pi(\mathcal{B}_3))$ is a tripartition of $\Pi$. Moreover, $\pi(\mathcal{I}_1) = \mathcal{D}_1$, $\pi(\mathcal{I}_2) = \mathcal{D}_2$, $\pi(\mathcal{I}_3) = \mathcal{D}_1 \cup \mathcal{D}_3$, $\pi(\mathcal{I}_4) = \mathcal{D}_2 \cup \mathcal{D}_3$. Then, by applying Corollary 1 we obtain that $(s_1, s_2, s_3, s_4)$ is a SOP. □

Traditionally the SOP can be constructed based on the fragmented SOP which requires only the contrariety and contradiction relations (which goes back to Aristotle's *De Interpretatione* 6–7, 17b.17–26, see [19, Sect. 2]). This result also holds in our framework:

**Theorem 2** *The quadruple $(s_1, s_2, s_3, s_4)$ of sentences, with $s_k : (\mathcal{F}, \mathcal{I}_k)$, $k = 1, 2, 3, 4$, is a SOP iff relations (a) and (c) in Definition 9 are satisfied.*

*Proof* (⇒) It follows directly from Definition 9. (⇐) We prove that (d) and (b) in Definition 9 follow from (a) and (c). If $\pi(\mathcal{I}_1) = \emptyset$, then of course $\pi(\mathcal{I}_1) \subseteq \pi(\mathcal{I}_3)$. If $\pi(\mathcal{I}_1) \neq \emptyset$, let $x \in \pi(\mathcal{I}_1) \subseteq \varPi$, from (a) it follows that $x \notin \pi(\mathcal{I}_2)$, and since (c) requires $\pi(\mathcal{I}_2) \cup \pi(\mathcal{I}_3) = \varPi$, we obtain $x \in \pi(\mathcal{I}_3)$. Thus, $\pi(\mathcal{I}_1) \subseteq \pi(\mathcal{I}_3)$; likewise, $\pi(\mathcal{I}_2) \subseteq \pi(\mathcal{I}_4)$. Therefore, (d) is satisfied. Now we prove that (b) is satisfied, i.e. $\pi(\mathcal{I}_3) \cup \pi(\mathcal{I}_4) = \varPi$. Of course, $\pi(\mathcal{I}_3) \cup \pi(\mathcal{I}_4) \subseteq \varPi$. Let $x \in \varPi$. If $x \notin \pi(\mathcal{I}_3)$, then, $x \in \pi(\mathcal{I}_2)$ from (c). Moreover, from (d), $x \in \pi(\mathcal{I}_4)$. Then, $\varPi \subseteq \pi(\mathcal{I}_3) \cup \pi(\mathcal{I}_4)$. Therefore, (b) is satisfied. □

*Remark 5* Given two sentences $s_1$ and $s_2$ that are contraries, then the quadruple $(s_1, s_2, \bar{s}_2, \bar{s}_1)$ is a SOP.

## 4 Square of Opposition and Generalized Quantifiers

Let $\mathcal{F}$ be a conditional event $P|S$ (where $S \neq \bot$) and $(\mathcal{B}_1(x), \mathcal{B}_2(x), \mathcal{B}_3(x))$ be a tripartition of $[0, 1]$, where $\mathcal{B}_1(x) = [x, 1]$, $\mathcal{B}_2(x) = [0, 1 - x]$, $\mathcal{B}_3 = ]1 - x, x[$ and $x \in ]\frac{1}{2}, 1]$. Consider the quadruple of sentences $(A(x), E(x), I(x), O(x))$, with $A(x) : (P|S, \mathcal{I}_{A(x)})$, $E(x) : (P|S, \mathcal{I}_{E(x)})$, $I(x) : (P|S, \mathcal{I}_{I(x)})$, $O(x) : (P|S, \mathcal{I}_{O(x)})$, where $\mathcal{I}_{A(x)} = \mathcal{B}_1 = [x, 1]$, $\mathcal{I}_{E(x)} = \mathcal{B}_2 = [0, 1 - x]$, $\mathcal{I}_{I(x)} = \mathcal{B}_1 \cup \mathcal{B}_3 = ]1 - x, 1]$, and $\mathcal{I}_{O(x)} = \mathcal{B}_2 \cup \mathcal{B}_3 = [0, x[$. By applying Corollary 2 with $(s_1, s_2, s_3, s_4) = (A(x), E(x), I(x), O(x))$, it follows that $(A(x), E(x), I(x), O(x))$ is a SOP for any $x \in ]\frac{1}{2}, 1]$ (see Fig. 1). We recall that in presence of some logical relations between $P$ and $S$ the set $\varPi$ could be a strict subset of $[0, 1]$. In particular, we have the following three cases (see, [15, 16]): (i) if $P \wedge S \neq \bot$ and $P \wedge S \neq S$, then $\varPi = [0, 1]$; (ii) if $P \wedge S = S$, then $\varPi = \{1\}$; (iii) if $P \wedge S = \bot$, then $\varPi = \{0\}$. The quadruple $(A(x), E(x), I(x), O(x))$, with the threshold $\frac{1}{2} < x \leq 1$, is a SOP in each of the three cases. In particular we obtain: case (i) $\pi(\mathcal{I}_{A(x)}) = \mathcal{I}_{A(x)}$, $\pi(\mathcal{I}_{E(x)}) = \mathcal{I}_{E(x)}$, $\pi(\mathcal{I}_{I(x)}) = \mathcal{I}_{I(x)}$, and $\pi(\mathcal{I}_{O(x)}) = \mathcal{I}_{O(x)}$; case (ii): $\pi(\mathcal{I}_{A(x)}) = \{1\}$,

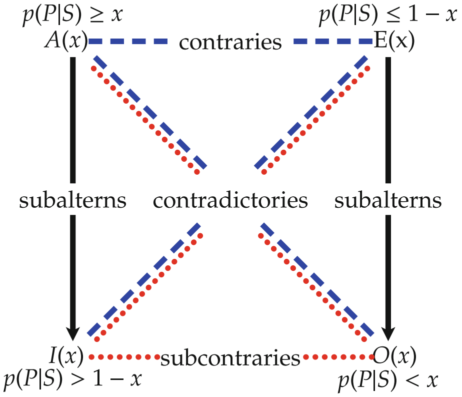


**Fig. 1** Probabilistic SOP defined on the four sentence types $(A(x), E(x), I(x), O(x))$ with the threshold $x \in ]\frac{1}{2}, 1]$ (see also Table 1). It provides a new interpretation of the traditional SOP (see, e.g., [19]), where the corners are labeled by “Every $S$ is $P$” (A), “No $S$ is $P$” (E), “Some $S$ is $P$” (I), and “Some $S$ is not $P$” (O)

**Table 1** Probabilistic interpretation of the sentence types $A$, $E$, $I$, and $O$ involving generalized quantifiers $Q$ defined by a threshold $x$ (with $x \in ]\frac{1}{2}, 1]$) on the subject $S$ and predicate $P$ and the respective imprecise probabilistic assessments $\mathcal{I}_{A(x)}, \mathcal{I}_{E(x)}, \mathcal{I}_{I(x)}$, and $\mathcal{I}_{O(x)}$ on the conditional event $P|S$ (above). When $x = 1$, we obtain our probabilistic interpretation of the traditional sentence types $A$, $E$, $I$, and $O$ (below)

| Sentence | | Probability constraints | Assessment on $P\|S$ |
|---|---|---|---|
| $A(x)$ : | ($Q_{\geq x}$ $S$ are $P$) | $p(P\|S) \geq x$ | $\mathcal{I}_{A(x)} = [x, 1]$ |
| $E(x)$ : | ($Q_{\geq x}$ $S$ are not $P$) | $p(\overline{P}\|S) \geq x$ | $\mathcal{I}_{E(x)} = [0, 1 - x]$ |
| $I(x)$ : | ($Q_{>1-x}$ $S$ are $P$) | $p(P\|S) > 1 - x$ | $\mathcal{I}_{I(x)} = ]1 - x, 1]$ |
| $O(x)$ : | ($Q_{>1-x}$ $S$ are not $P$) | $p(\overline{P}\|S) > 1 - x$ | $\mathcal{I}_{O(x)} = [0, x[$ |
| $A(1)$ : | (Every $S$ is $P$) | $p(P\|S) = 1$ | $\mathcal{I}_A = [1, 1]$ |
| $E(1)$ : | (No $S$ is $P$) | $p(\overline{P}\|S) = 1$ | $\mathcal{I}_E = [0, 0]$ |
| $I(1)$ : | (Some $S$ is $P$) | $p(P\|S) > 0$ | $\mathcal{I}_I = ]0, 1]$ |
| $O(1)$ : | (Some $S$ is not $P$) | $p(\overline{P}\|S) > 0$ | $\mathcal{I}_O = [0, 1[$ |

$\pi(\mathcal{I}_{E(x)}) = \emptyset$,$\pi(\mathcal{I}_{I(x)}) = \{1\}$, and $\pi(\mathcal{I}_{O(x)}) = \emptyset$; case (iii): $\pi(\mathcal{I}_{A(x)}) = \emptyset$, $\pi(\mathcal{I}_{E(x)}) = \{1\}$,$\pi(\mathcal{I}_{I(x)}) = \emptyset$, and $\pi(\mathcal{I}_{O(x)}) = \{1\}$. We note that in cases (ii) and (iii) we obtain degenerated squares each, where—apart from the contradictory relations—all relations are strengthened. Specifically, both contrary and the subcontrary become contradictory relations. Moreover, both subalternation relations become symmetric. As by coherence $p(P|S) + p(\overline{P}|S) = 1$, a sentence $s : (P|S, \mathcal{I})$ is equivalent to the sentence $s' : (\overline{P}|S, \overline{\mathcal{I}})$, where $\overline{\mathcal{I}} = [0, 1] \setminus \mathcal{I}$. Table 1 presents generalization of basic sentence types $A(x)$, $E(x)$, $I(x)$, and $O(x)$ involving generalized quantifiers $Q$. The generalized quantifiers are defined on a threshold $x > \frac{1}{2}$. The value of the threshold may be context dependent and provides lots of flexibility for modeling various instances of generalized quantifiers (like “most”, “almost all”). In the extreme case $x = 1$ we obtain the probabilistic interpretation under coherence of the basic sentence types involved in the traditional SOP $(A, E, I, O)$ (see [13, 14]).

## 5 Concluding Remarks

Based on tools developed in Sect. 3, we can construct a *hexagon of opposition* by starting from a square (see, e.g., [5]). More precisely, given a SOP $(s_1, s_2, s_3, s_4)$, by setting $A = s_1$, $E = s_2$, $I = s_3$, $O = s_4$, $U = s_1 \vee s_2$, $Y = s_3 \wedge s_4$, the tuple $(A, E, I, O, U, Y)$ defines a hexagon of opposition, which we will elaborate in another paper. Moreover, we note the square presented in Sect. 4 can serve as a new rationality framework for investigating generalized quantifiers, which are psychologically much more plausible compared to the traditional logical quantifiers, as the latter are either too strict (∀) or too weak (∃) for formalizing everyday life sentences (see [18, 20, 21]).

**Acknowledgments** We thank A. Gilio and the anonymous referees for useful comments. We thank Villa Vigoni (*Human Rationality: Probabilistic Points of View*). N. Pfeifer is supported by his DFG project PF 740/2-2 (within the SPP1516). G. Sanfilippo is supported by the INdAM–GNAMPA Projects (2016 Grant U 2016/000391 and 2015 grant U 2015/000418).

## References

1. Béziau J-Y, Read S (2014) Editorial: Square of opposition: a diagram and a theory in historical perspective. Hist Philos Log 35(4):315–316
2. Biazzo V, Gilio A (2000) A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments. IJAR 24(2–3):251–272
3. Biazzo V, Gilio A, Lukasiewicz T, Sanfilippo G (2005) Probabilistic logic under coherence: complexity and algorithms. AMAI 45(1–2):35–81
4. Capotorti A, Lad F, Sanfilippo G (2007) Reassessing accuracy rates of median decisions. Am Stat 61(2):132–138
5. Ciucci D, Dubois D, Prade H (2015) Structures of opposition induced by relations. Ann Math Artif Intell, pp 1–23
6. Coletti G, Petturiti D, Vantaggi B (2014) Possibilistic and probabilistic likelihood functions and their extensions: common features and specific characteristics. Fuzzy Sets Syst 250:25–51
7. Coletti G, Scozzafava R (2002) Probabilistic logic in a coherent setting. Kluwer
8. Coletti G, Scozzafava R, Vantaggi B (2015) Possibilistic and probabilistic logic under coherence: default reasoning and System P. Mathematica Slovaca 65(4):863–890
9. Dubois D, Prade H (2012) From Blanché's hexagonal organization of concepts to formal concept analysis and possibility theory. Logica Universalis 6:149–169
10. Gilio A (2002) Probabilistic reasoning under coherence in System P. AMAI 34:5–34
11. Gilio A, Ingrassia S (1998) Totally coherent set-valued probability assessments. Kybernetika 34(1):3–15
12. Gilio A, Over DE, Pfeifer N, Sanfilippo G, Centering and compound conditionals under coherence. In this issue
13. Gilio A, Pfeifer N, Sanfilippo G (2015) Transitive reasoning with imprecise probabilities. In: ECSQARU'15, vol 9161 of LNAI, Springer, Berlin, pp 95–105
14. Gilio A, Pfeifer N, Sanfilippo G (2016) Transitivity in coherence-based probability logic. J Appl Log 14:46–64
15. Gilio A, Sanfilippo G (2013) Probabilistic entailment in the setting of coherence: the role of quasi conjunction and inclusion relation. IJAR 54(4):513–525
16. Gilio A, Sanfilippo G (2013) Quasi conjunction, quasi disjunction, t-norms and t-conorms: probabilistic aspects. Inf Sci 245:146–167
17. Gilio A, Sanfilippo G (2014) Conditional random quantities and compounds of conditionals. Studia Logica 102(4):709–729
18. Oaksford M, Chater N (2007) Bayesian rationality. OUP, Oxford
19. Parsons T (2015) The traditional square of opposition. In Zalta EN, (ed), The Stanford Encyclopedia of Philosophy. Summer 2015 edition
20. Pfeifer N (2006) Contemporary syllogistics: comparative and quantitative syllogisms. In Argumentation in Theorie und Praxis, pp 57–71. LIT Verlag, Wien
21. Pfeifer N (2013) The new psychology of reasoning: a mental probability logical perspective. Thinking Reasoning 19(3–4):329–345
22. Pfeifer N (2014) Reasoning about uncertain conditionals. Studia Logica 102(4):849–866
23. Pfeifer N, Kleiter GD (2009) Framing human inference by coherence based probability logic. J Appl Log 7(2):206–217

# Testing of Coarsening Mechanisms: Coarsening at Random Versus Subgroup Independence

**Julia Plass, Marco E.G.V. Cattaneo, Georg Schollmeyer and Thomas Augustin**

**Abstract** Since coarse(ned) data naturally induce set-valued estimators, analysts often assume coarsening at random (CAR) to force them to be single-valued. Using the PASS data as an example, we re-illustrate the impossibility to test CAR and contrast it to another type of uninformative coarsening called subgroup independence (SI). It turns out that SI is testable here.

**Keywords** Coarse data · Missing data · Coarsening at random (CAR) · Hypothesis testing · Likelihood-ratio test

## 1 The Problem of Testing Coarsening Mechanisms

Traditional statistical methods dealing with missing data (e.g. EM algorithm, imputation techniques) require identifiability of parameters, which frequently tempts analysts to make the *missing at random* (MAR) assumption [8] simply for pragmatic reasons without justifications in substance (e.g. [6]). Since MAR is not testable (e.g. [9]), this way to proceed is especially alarming.

Looking at the problem in a more general way, incomplete observations may be included not only in the sense of missing, but also coarse(ned) data. In this way, additionally to fully observed and unobserved, also partially observed values are

J. Plass (✉) · G. Schollmeyer · T. Augustin
Department of Statistics, LMU Munich, Munich, Germany
e-mail: julia.plass@stat.uni-muenchen.de

G. Schollmeyer
e-mail: georg.schollmeyer@stat.uni-muenchen.de

T. Augustin
e-mail: augustin@stat.uni-muenchen.de

M. Cattaneo
Department of Mathematics, University of Hull, Kingston upon Hull, UK
e-mail: m.cattaneo@hull.ac.uk

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_51

considered.[1] In the context of coarse data, the *coarsening at random* (CAR) assumption (e.g. [5]) is the analogue of MAR. Although the impossibility of testing CAR is already known from literature, providing an intuitive insight into this point will be a first goal of this paper. Apart from CAR, we focus on another, in a sense dual, assumption that we called *subgroup independence* (SI) in [11]. In our categorical setting (cf. Sect. 2), SI not only makes parameters identifiable, but is also testable as demonstrated here. Thus, we elaborate the substantial difference in the testability of CAR and SI and start with illustrating both assumptions by a running example based on the PASS data in Sect. 2 [14]. In Sect. 3 we sketch the crucial argument of the estimation and show how the generally set-valued estimators are refined by implying CAR or SI. Testability of both assumptions is discussed in Sect. 4, where a likelihood-ratio test is suggested for SI.

## 2 Coarsening Models: CAR and SI

Throughout this paper, we refer to the case of a coarse categorical response variable $Y$ and one precisely observed binary covariate $X$. The results may be easily transferred to cases with more than one arbitrary categorical covariates by using dummy variables and conditioning on the then emerged subgroups. For sake of conciseness, the example refers to the case of a binary $Y$, where coarsening corresponds to missingness, but all results are applicable in a general categorical setting.

We approach the problem of coarse data in our setting by distinguishing between a latent and an observed world: A random sample of a categorical response variable $Y_1, \dots, Y_n$ with realizations $y_1, \dots, y_n$ in sample space $\Omega_Y$ is part of the latent world. The basic goal consists of estimating the individual probabilities $\pi_{xy} = P(Y_i = y | X_i = x)$ given the precise values of a categorical covariate $X$ with sample space $\Omega_X$. Unfavorably, the values of $Y$ can only be observed partially and thus the realizations $\mathfrak{y}_1, \dots, \mathfrak{y}_n$ of a sample $\mathcal{Y}_1, \dots, \mathcal{Y}_n$ of a random object $\mathcal{Y}$ within sample space $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \emptyset$ constitute the observed world, with $\mathfrak{y}_i \ni y_i$.[2] A connection between both worlds, and thus between $\pi_{xy}$ and $p_{x\mathfrak{y}} = P(\mathcal{Y}_i = \mathfrak{y} | X_i = x)$, is established via an observation model governed by the coarsening parameters $q_{\mathfrak{y}|xy} = P(\mathcal{Y}_i = \mathfrak{y} | X_i = x, Y_i = y)$ with $\mathfrak{y} \in \Omega_{\mathcal{Y}}$, $y \in \Omega_Y$, and $x \in \Omega_X$. As the dimension of these coarsening parameters increases considerably with $|\Omega_X|$ and $|\Omega_Y|$, for reasons of conciseness, we mainly confine ourselves to the discussion of the example with $\Omega_X = \{0, \ 1\}$, $\Omega_Y = \{a, \ b\}$, and thus $\Omega_{\mathcal{Y}} = \{\{a\}, \ \{b\}, \ \{a, b\}\}$, where "$\{a, b\}$" denotes the only coarse observation, which corresponds to a missing one in this case. Assuming only error-freeness, we generally refrain from making strict assumptions on $q_{\mathfrak{y}|xy}$. In contrast to this, under CAR and SI the coarsening parameters are strongly restricted.

[1] When dealing with coarse data, it is important to distinguish between epistemic data imprecision, considered here, and ontic data imprecision (cf. [2]).

[2] This error-freeness implies that $Y$ is an almost sure selector of $\mathcal{Y}$ (in the sense of e.g. [10]).

**Table 1** Data of the PASS example

| UBII | Income | Observed counts | Total counts |
|---|---|---|---|
| 0 | $\{a\}$ | $n_{0\{a\}} = 38$ | $n_0 = 518$ |
| | $\{b\}$ | $n_{0\{b\}} = 385$ | |
| | $\{a, b\}$ | $n_{0\{a,b\}} = 95$ | |
| 1 | $\{a\}$ | $n_{1\{a\}} = 36$ | $n_1 = 87$ |
| | $\{b\}$ | $n_{1\{b\}} = 42$ | |
| | $\{a, b\}$ | $n_{1\{a,b\}} = 9$ | |

Heitjan and Rubin [4] consider maximum likelihood estimation in coarse data situations by deriving assumptions simplifying the likelihood. These assumptions—CAR and distinct parameters—make the coarsening *ignorable* (e.g. [8]). The CAR assumption requires constant coarsening parameters $q_{\mathfrak{y}|xy}$, regardless which true value $y$ is underlying subject to the condition that it matches with the fixed observed value $\mathfrak{y}$. The strong limitation of this assumption is illustrated by the running example generally introduced in the following box.

**Running example** (Table 1 shows the summary of the data)

- German Panel Study "Labour Market and Social Security" [14] (PASS, wave 5, 2011)
- $Y$: income $<$ 1000€ (a) or $\geq$ 1000€ (b) $\Rightarrow y \in \{a,\ b\}$
- $\mathcal{Y}$: some respondents give no suitable answer ($\{a, b\}$: $y = a$ or $y = b$) $\Rightarrow \mathfrak{y} \in \{\{a\}, \{b\}, \{a, b\}\} \Rightarrow$ coarse answer $\{a, b\}$ is missing observation
- $X$: receipt of Unemployment Benefit II (UBII), $x \in \{0 \text{ (no)},\ 1 \text{ (yes)}\}$

Referring to the example, under CAR, which coincides here with MAR,[3] the probability of giving no suitable answer is taken to be independent of the true income category in both subgroups split by UBII, i.e.

$$q_{\{a,b\}|0a} = q_{\{a,b\}|0b} \text{ and } q_{\{a,b\}|1a} = q_{\{a,b\}|1b}.$$

Generally, CAR could be quite problematic in this context, as practical experiences show that reporting missing or coarsened answers is notably common in specific income groups (e.g. [7]).

If the data are missing not at random (MNAR) [8], commonly the missingness process is modelled by including parametric assumptions (e.g. [4, 8]) or a cautious

[3] The PASS data provide income in different levels of coarseness induced by follow-up questions for non-respondents. For sake of simplicity, we consider only the income question explained in the box, but the study provides also coarse ordinal data in the general sense.

procedure is chosen ending up in set-valued estimators (cf. e.g. [3, 11, 17]). For the categorical case, there is a special case of MNAR, in which single-valued estimators are obtained without parametric assumptions. For motivating this case, one can further differentiate MNAR, distinguishing between the situation where missingness depends on both the values of the response $Y$ and the covariate $X$ and the situation where it depends on the values of $Y$ only. Referring to the related coarsening case, the latter case corresponds to SI investigated in [11]. This independence from the covariate value shows, beside CAR, an alternative kind of uninformative coarsening.

Again, one should use this assumption cautiously: Under SI, giving a coarse answer is taken to be independent of the UBII given the value of $Y$, i.e.

$$q_{\{a,b\}|0a} = q_{\{a,b\}|1a} \text{ and } q_{\{a,b\}|0b} = q_{\{a,b\}|1b}.$$

Mostly, this turns out to be doubtful, as the receipt of the UBII influences the income, which typically has an impact on the non-response to the income question.

## 3 Estimation: General Approach, CAR and SI

This section recalls some important aspects of an approach developed in [11] by sketching the basic idea of the therein considered cautious, likelihood-based estimation technique. The resulting estimators are not only given for the general case, but also when the assumptions in focus are included.

To estimate $(\pi_{xy})_{x\in\Omega_X, y\in\Omega_Y}$ of the latent world, basically three steps are accomplished. Firstly, we determine the maximum likelihood estimator (MLE) $(\hat{p}_{x\mathfrak{y}})_{x\in\Omega_X, \mathfrak{y}\in\Omega_{\mathcal{Y}}}$ in the observed world. Since the counts $(n_{x\mathfrak{y}})_{x\in\Omega_X, \mathfrak{y}\in\Omega_{\mathcal{Y}}}$ are multinomially distributed, the unique MLE is obtained by the relative frequencies of the respective categories, coarse categories treated as own categories. Secondly, we connect the parameters of both worlds by a mapping $\Phi$. For the binary case with $x \in \{0, 1\}$ one obtains $\Phi : [0, 1]^6 \to [0, 1]^4$ with

$$\Phi \begin{pmatrix} \pi_{xa} \\ q_{\{a,b\}|xa} \\ q_{\{a,b\}|xb} \end{pmatrix} = \begin{pmatrix} \pi_{xa} \cdot (1 - q_{\{a,b\}|xa}) \\ (1 - \pi_{xa}) \cdot (1 - q_{\{a,b\}|xb}) \end{pmatrix} = \begin{pmatrix} p_{x\{a\}} \\ p_{x\{b\}} \end{pmatrix}. \tag{1}$$

Thirdly, by the invariance of the likelihood under parameter transformations, we may incorporate the parametrization in terms of $\pi_{xy}$ and $q_{\mathfrak{y}|xy}$ into the likelihood of the observed world. Since the mapping $\Phi$ is generally not injective, we obtain set-valued estimators $\hat{\pi}_{xy}$ and $\hat{q}_{\mathfrak{y}|xy}$, namely

$$\hat{\pi}_{xa} \in \left[\frac{n_{x\{a\}}}{n_x}, \frac{n_{x\{a\}} + n_{x\{a,b\}}}{n_x}\right], \quad \hat{q}_{\{a,b\}|xy} \in \left[0, \frac{n_{x\{a,b\}}}{n_{x\{y\}} + n_{x\{a,b\}}}\right], \tag{2}$$

with $x \in \{0,\ 1\}$ and $y \in \{a,\ b\}$. Points in these sets are constrained by the relationships in $\Phi$. In the spirit of the methodology of *partial identification* [9], these sets may be refined by including assumptions about the coarsening justified from the application standpoint. Very strict assumptions may induce point identified parameters, as estimation under CAR or SI in the categorical case shows.[4]

Including CAR, i.e. restricting the set of possible coarsening mechanisms to $q_{\{a,b\}|xa} = q_{\{a,b\}|xb}$ with $x \in \{0,\ 1\}$, induces an injective mapping $\Phi$ leading to the point-valued estimators

$$\hat{\pi}_{xa}^{CAR} = \frac{n_{x\{a\}}}{n_{x\{a\}} + n_{x\{b\}}}, \quad \hat{q}_{\{a,b\}|xa}^{CAR} = \hat{q}_{\{a,b\}|xb}^{CAR} = \frac{n_{x\{a,b\}}}{n_x}. \tag{3}$$

Thus, under this type of uninformative coarsening, $\hat{\pi}_{xa}$ corresponds here to the proportion of $\{a\}$-observations in subgroup $x$ ignoring all coarse values and $\hat{q}_{\{a,b\}|xa} = \hat{q}_{\{a,b\}|xb}$ is the proportion of observed $\{a, b\}$ in subgroup $x$.

Under rather weak regularity conditions, namely $\pi_{0a} \neq \pi_{1a}$, $\pi_{0a} \notin \{0, 1\}$, and $\pi_{1a} \notin \{0, 1\}$ for $x \in \{0, 1\}$, also under SI the mapping $\Phi$ becomes injective (cf. [12]) in our categorical setting. Hence, point-valued estimators

$$\begin{aligned} \hat{\pi}_{xa}^{SI} &= \frac{n_{x\{a\}}}{n_x} \, \frac{n_0\, n_{1\{b\}} - n_{0\{b\}}\, n_1}{n_{0\{a\}}\, n_{1\{b\}} - n_{0\{b\}}\, n_{1\{a\}}}, \\ \hat{q}_{\{a,b\}|xa}^{SI} &= \frac{n_{0\{a,b\}}\, n_{1\{b\}} - n_{0\{b\}}\, n_{1\{a,b\}}}{n_0\, n_{1\{b\}} - n_{0\{b\}}\, n_1}, \\ \hat{q}_{\{a,b\}|xb}^{SI} &= \frac{n_{0\{a,b\}}\, n_{1\{a\}} - n_{0\{a\}}\, n_{1\{a,b\}}}{n_0\, n_{1\{a\}} - n_{0\{a\}}\, n_1} \end{aligned} \tag{4}$$

are obtained, provided they are well-defined and inside [0, 1].

## 4 Testing

Due to the substantial bias of $\hat{\pi}_{xy}$ if CAR or SI are wrongly assumed (cf. e.g. [12]), testing these assumptions is of particular interest. Although it is already established that it is not possible to test whether the CAR condition holds (e.g. [9]), it may be insightful, in particular in the light of Sect. 4.2, to address this impossibility in the context of the example.

[4] Identifiability may not only be obtained by assumptions on the coarsening: e.g. for discrete graphical models with one hidden node, conditions based on the associated concentration graph are used in [13].

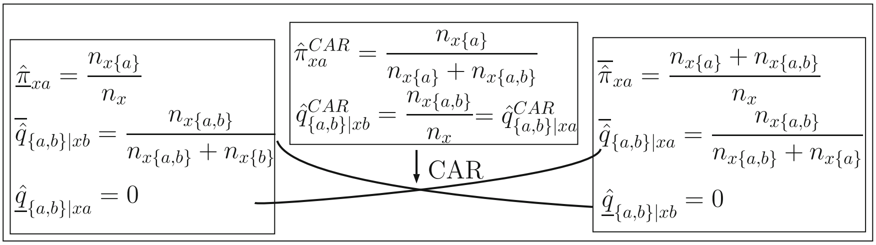

$\underline{\hat{\pi}}_{xa} = \frac{n_{x\{a\}}}{n_x}$
$\overline{\hat{q}}_{\{a,b\}|xb} = \frac{n_{x\{a,b\}}}{n_{x\{a,b\}} + n_{x\{b\}}}$
$\underline{\hat{q}}_{\{a,b\}|xa} = 0$
$\hat{\pi}^{CAR}_{xa} = \frac{n_{x\{a\}}}{n_{x\{a\}} + n_{x\{a,b\}}}$
$\hat{q}^{CAR}_{\{a,b\}|xb} = \frac{n_{x\{a,b\}}}{n_x} = \hat{q}^{CAR}_{\{a,b\}|xa}$
CAR
$\overline{\hat{\pi}}_{xa} = \frac{n_{x\{a\}} + n_{x\{a,b\}}}{n_x}$
$\overline{\hat{q}}_{\{a,b\}|xa} = \frac{n_{x\{a,b\}}}{n_{x\{a,b\}} + n_{x\{a\}}}$
$\underline{\hat{q}}_{\{a,b\}|xb} = 0$

**Fig. 1** Since the relationships expressed via $\varPhi$ in (1) have to be met, only specific points from the set-valued estimators in (2) are combinable, ranging from ($\underline{\hat{\pi}}_{xa}$, $\underline{\hat{q}}_{\{a,b\}|xa}$, $\overline{\hat{q}}_{\{a,b\}|xb}$) to ($\overline{\hat{\pi}}_{xa}$, $\overline{\hat{q}}_{\{a,b\}|xa}$, $\underline{\hat{q}}_{\{a,b\}|xb}$) with the CAR case always included

## 4.1 Testing of CAR

A closer consideration of (3) already indicates that CAR can never be rejected without including additional assumptions about the coarsening. This point is illustrated in Fig. 1 by showing the interaction between points in the intervals in (2). Thus, this uninformative coarsening—in the sense that all coarse observations are ignored—is always a possible scenario included in the general set-valued estimators in (2).

Exemplary for subgroup 0, under CAR we obtain $\hat{\pi}^{CAR}_{0a} = 0.09$, $\hat{q}^{CAR}_{\{a,b\}|0a} = \hat{q}^{CAR}_{\{a,b\}|0b} = 0.18$, which may not be excluded from the general estimators $\hat{\pi}_{0a} \in [0.07,\ 0.26]$, $\hat{q}_{\{a,b\}|0a} \in [0, 0.71]$ and $\hat{q}_{\{a,b\}|0b} \in [0, 0.20]$ unless further assumptions as e.g. "respondents from the high income group tend to give coarse answers more likely" are justified.

Nevertheless, there are several approaches that show how testability of CAR is achieved by distributional assumptions (e.g. [5]), e.g. the naive Bayes assumption [6], or by the inclusion of instrumental variables (cf. [1]).

## 4.2 Testing of SI

Applying the estimators in (4) to the example, one obtains $\hat{\pi}^{SI}_{0a} = 0.42$, $\hat{\pi}^{SI}_{1a} = 0.40$, $\hat{q}^{SI}_{\{a,b\}|0a} = \hat{q}^{SI}_{\{a,b\}|1a} = -0.04$, and $\hat{q}^{SI}_{\{a,b\}|0b} = \hat{q}^{SI}_{\{a,b\}|1b} = 0.20$ partly outside $[0, 1]$. This shows that there are data situations that might hint to (partial) incompatibility with SI. In general for the categorical case, a statistical test for the following hypotheses can be constructed:

$$H_0 : q_{\mathbf{y}|xy} = q_{\mathbf{y}|x'y} \text{ for all } \mathbf{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y,$$
$$H_1 : q_{\mathbf{y}|xy} \neq q_{\mathbf{y}|x'y} \text{ for some } \mathbf{y} \in \Omega_{\mathcal{Y}},\ x, x' \in \Omega_X,\ y \in \Omega_Y.$$

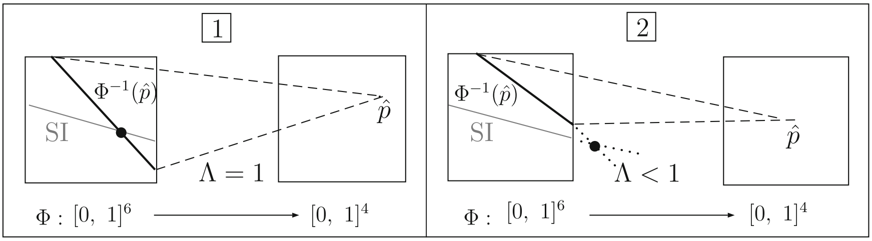


**Fig. 2** The impact on $\Lambda$ of two substantially differing data situations is illustrated

As test statistic we can use the likelihood ratio (e.g. [16])

$$\Lambda(\mathfrak{y}_1, \dots, \mathfrak{y}_n, x_1, \dots, x_n) = \frac{\sup_{H_0} L(\vartheta || \mathfrak{y}_1, \dots, \mathfrak{y}_n, x_1, \dots, x_n)}{\sup_{H_0 \cup H_1} L(\vartheta || \mathfrak{y}_1, \dots, \mathfrak{y}_n, x_1, \dots, x_n)},$$

here with $\boldsymbol{\vartheta} = (\pi_{0a}, \pi_{1a}, q_{\{a,b\}|0a}, q_{\{a,b\}|1a}, q_{\{a,b\}|0b}, q_{\{a,b\}|1b})^T$.[5] In fact, recent simulation studies corroborate the decrease of $\Lambda$ with deviation from SI (cf. [12]). The sensitivity of $\Lambda$ with regard to the test considered here is also illustrated informally in Fig. 2 by depicting $\boldsymbol{\Phi}$ in (1) for two data situations, where only the second one gives evidence against SI. The gray line symbolizes all arguments satisfying SI, while the bold line represents all arguments maximizing the likelihood (i.e. all values in (2) compatible with each other). The intersection of both lines represents the values in (4), and if it is included in the domain of $\boldsymbol{\Phi}$ (cf. first case of Fig. 2), the same maximal value of the likelihood is obtained regardless of including SI or not, resulting in $\Lambda = 1$. An intersection outside the domain (cf. second case of Fig. 2) induces a lower value of the likelihood under SI, also reflected in $\Lambda < 1$. For the example one obtains $\Lambda \approx 0.93$ and thus there is a slight evidence against SI based on a direct interpretation of the likelihood ratio, while setting a general decision rule depending on a significance level $\alpha$ remains as an open problem.

## 5 Conclusion

We focused on the testability of CAR and SI by investigating the compatibility of the estimators (3) and (4) with the observed data. While CAR is generally not testable, SI may be tested and a "pure likelihood" approach was proposed. To obtain a statistical test for SI at a fixed level of significance $\alpha$, we want to determine the (asymptotic) distribution of $-2 \log \Lambda$ under $H_0$ next, which is expected to deviate from the $\chi^2$-distribution of the standard case. Furthermore, a generalized version of SI—in the

[5] While the denominator of $\Lambda$ can be obtained using any values in (2) compatible with each other, the numerator must in general be calculated by numerical optimization. Alternatives to this statistic include a test decision based on uncertainty regions [15].

sense of assuming particular coarsening parameters to be known multiples of each other—will allow for a more flexible application of this hypothesis test.

**Acknowledgments** We thank the Research Data Center at the Institute for Employment Research, Nuremberg, especially Mark Trappmann and Anja Wurdack, for the access to the PASS data and their support in practical matters. Furthermore, we are grateful to two anonymous reviewers for their very helpful remarks and valuable suggestions.

## References

1. Breunig C (2015) Testing missing at random using instrumental variables. Humboldt University, Collaborative Research Center 649. https://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2015-016.pdf
2. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: ontic vs. epistemic views. Int J Approximate Reasoning 55:1502–1518
3. Denoeux T (2014) Likelihood-based belief function: justification and some extensions to low-quality data. Int J Approximate Reasoning 55:1535–1547
4. Heitjan D, Rubin D (1991) Ignorability and coarse data. Ann Stat 19:2244–2253
5. Jaeger M (2005) Ignorability for categorical data. Ann Stat 33:1964–1981
6. Jaeger M (2006) On testing the missing at random assumption. In: Fürnkranz J, Scheffer T, Spiliopoulou M (eds) Machine learning: ECML 2006. Springer
7. Korinek A, Mistiaen J, Ravallion M (2006) Survey nonresponse and the distribution of income. J Econ Inequal 4:33–55
8. Little R, Rubin D (2002) Statistical analysis with missing data, 2nd edn. Wiley
9. Manski C (2003) Partial identification of probability distributions. Springer
10. Nguyen H (2006) An introduction to random sets. CRC Press
11. Plass J, Augustin T, Cattaneo M, Schollmeyer G (2015) Statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression for coarse categorical data. In: Augustin T, Doria S, Miranda E, Quaeghebeur E (eds) ISIPTA '15. SIPTA
12. Plass J, Augustin T, Cattaneo M, Schollmeyer G (2016) Statistical modelling under epistemic data imprecision, LMU Munich. http://www.statistik.lmu.de/~jplass/forschung.html
13. Stanghellini E, Vantaggi B (2013) Identification of discrete concentration graph models with one hidden binary variable. Bernoulli 19:1920–1937
14. Trappmann M, Gundert S, Wenzig C, Gebhardt D (2010) PASS: a household panel survey for research on unemployment and poverty. Schmollers Jahrb 130:609–623
15. Vansteelandt S, Goetghebeur E, Kenward M, Molenberghs G (2006) Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Stat Sin 16:953–979
16. Wilks S (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Stat 9:60–62
17. Zaffalon M, Miranda E (2009) Conservative inference rule for uncertain reasoning under incompleteness. J Artif Intell Res 34:757–821

# Two-Sample Similarity Test for the Expected Value of Random Intervals

**Ana B. Ramos-Guajardo and Ángela Blanco-Fernández**

**Abstract** The similarity degree between the expectation of two random intervals is studied by means of a hypothesis testing procedure. For this purpose, a similarity measure for intervals is introduced based on the so-called Jaccard index for convex sets. The measure ranges from 0 (if both intervals are not similar at all, i.e., if they are not overlapped) to 1 (if both intervals are equal). A test statistic is proposed and its limit distribution is analyzed by considering asymptotic and bootstrap techniques. Some simulation studies are carried out to examine the behaviour of the approach.

## 1 Introduction

Interval data derive from experimental studies involving ranges, fluctuations, subjective perceptions, censored data, grouped data, and so on [5, 6, 9]. Random intervals (RIs) have been shown to model and handle suitably such kind of data in different settings [2, 3, 10, 11].

The Aumman expectation of a RI is also an interval and inferences concerning the Aumann expectation and, especially, hypothesis tests for the expected value of random intervals have been previously developed in the literature [4, 8]. Additionally, tests relaxing strict equalities have been also carried out as, for instance, inclusion tests for the Aumann expectation of RIs [12], or similarity tests for the expected value of an RI and a previously fixed interval [13].

The aim of this work is to develop a two-sample test for the similarity of the expectations of two RIs. The similarity measure to be considered is based on the classical Jaccard similarity coefficient for classical convex sets [7], which can be seen as a ratio of the Lebesgue measure of the intersection interval and the

A.B. Ramos-Guajardo (✉) · Á. Blanco-Fernández
Department of Statistics and Operational Research, University of Oviedo,
C/Calvo Sotelo, s/n, 33007 Oviedo, Spain
e-mail: ramosana@uniovi.es

Á. Blanco-Fernández
e-mail: blancoangela@uniovi.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_52

Lebesgue measure of the union interval [14]. A statistic to solve the test is introduced, and its asymptotic and bootstrap limit distributions are theoretically analyzed. The development of bootstrap techniques allows us to approximate the sampling distribution of the statistic in practice, since the asymptotic one depends on unknown parameters in general. Finally, simulation studies are developed to show the empirical behaviour of the procedure.

## 2 Preliminary Concepts

From now on, let $\mathcal{K}_c(\mathbb{R})$ denote the family of non-empty closed and bounded intervals in $\mathbb{R}$. An interval $A \in \mathcal{K}_c(\mathbb{R})$ can be characterized by either its (mid, spr) representation (i.e., $A = [\text{mid}A \pm \text{spr}A]$, with $\text{mid}A \in \mathbb{R}$ the mid-point or centre and $\text{spr}A \geq 0$ the spread or radius of $A$) or its (inf, sup) representation (i.e., $A = [\inf A, \sup A]$).

The usual interval arithmetic is based on the Minkowski's addition and the product by a scalar. It is expressed in terms of the (mid, spr) representation as $A_1 + \lambda A_2 = [(\text{mid}A_1 + \lambda \text{mid}A_2) \pm (\text{spr}A_1 + |\lambda|\text{spr}A_2)]$, for $A_1, A_2 \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$.

The Lebesgue measure of $A \in \mathcal{K}_c(\mathbb{R})$ is given by $\lambda(A) = 2\text{spr}A$. Obviously, the Lebesgue measure of the empty set is $\lambda(\emptyset) = 0$. In addition, the Lebesgue measure of the intersection between A and B, $\lambda(A \cap B)$, for any $A, B \in \mathcal{K}_c(\mathbb{R})$ can be expressed as follows [14]:

$$\max\left\{0, \min\left\{2\text{spr}A, 2\text{spr}B, \text{spr}A + \text{spr}B - |\text{mid}A - \text{mid}B|\right\}\right\}. \tag{1}$$

A measure of the degree of similarity between two intervals $A, B \in \mathcal{K}_c(\mathbb{R})$ can be defined according to the Jaccard coefficient [7] as

$$S(A, B) = \frac{\lambda(A \cap B)}{\lambda(A \cup B)}. \tag{2}$$

This similarity measure fulfils that $S(A, B) = 0$ iff $A \cap B = \emptyset$, $S(A, B) = 1$ iff $A = B$, and $S(A, B) \in (0, 1)$ iff $A \cap B \neq \emptyset$ and $A \neq B$. As an example, the similarity measure of two intervals $A$ and $B$ is $1/2$ whenever both intervals are overlapped and the length of $A$ is the double than the length of $B$, or viceversa.

Random variables modelling those situations in which intervals on $\mathcal{K}_c(\mathbb{R})$ are provided as outcomes are called *random intervals* (RIs). Given a probability space $(\Omega, \mathcal{A}, P)$, an RI is a Borel measurable mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ w.r.t. the well-known Hausdorff metric on $\mathcal{K}_c(\mathbb{R})$ [10]. It is equivalently shown that $X$ is an RI if both $\text{mid}X, \text{spr}X : \Omega \to \mathbb{R}$ are real-valued random variables and $\text{spr}X \geq 0$ a.s.-$[P]$.

Whenever $\text{mid}X, \text{spr}X \in L^1(\Omega, \mathcal{A}, P)$, it is possible to define the Aumann expected value of $X$ [1]. In terms of classical expectations it is expressed as $E([\text{mid}X \pm \text{spr}X]) = [E(\text{mid}X) \pm E(\text{spr}X)]$. Let $\{X_i\}_{i=1}^n$ be a simple random sample of $X$. The corresponding sample expectation of $X$ is defined coherently in terms of the interval arithmetic as $\overline{X} = (1/n)\sum_{i=1}^n X_i$, and it fulfils $\overline{X} = [\overline{\text{mid}X} \pm \overline{\text{spr}X}]$.

## 3 Similarity Test for the Expected Values of Two RIs

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X, Y : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ be two RIs such that $\text{spr}E(X) > 0$ and $\text{spr}E(Y) > 0$. Some mild conditions are assumed to guarantee the existence of the involved moments and to avoid trivial situations (as, for instance, the singularity of the covariance matrix). Thus, $X$ and $Y$ are supposed to belong to the following class of random intervals:

$$\mathcal{P} = \Big\{X : \Omega \to \mathcal{K}_c(\mathbb{R}) \mid \sigma^2_{\text{mid}X} < \infty, 0 < \sigma^2_{\text{spr}X} < \infty \\ \wedge (Cov(\text{mid}X,\ \text{spr}X))^2 \neq \sigma^2_{\text{mid}X}\sigma^2_{\text{spr}X}\Big\}.$$

Given $d \in [0, 1]$, the aim is to test

$$H_0 : S(E(X), E(Y)) \geq d \text{ versus } H_1 : S(E(X), E(Y)) < d. \tag{3}$$

The alternative one-sided and two-sided tests (that is, those analyzing if the Jaccard index of the expectations equals $d$ or if it is greater than or equal to $d$) could be analogously studied. We focus our attention in (3) since it seems to be the most appealing for practical applications. From (1) and (2) it is straightforward to show that the null hypothesis of the test (3) can be equivalently expressed as

$$H_0 : \max \Big\{d\,\text{spr}E(Y) - \text{spr}E(X), d\,\text{spr}E(X) - \text{spr}E(Y), \\ (1+d)\,|\text{mid}E(X) - \text{mid}E(Y)| \\ + (d-1)\,(\text{spr}E(X) + \text{spr}E(Y))\Big\} \leq 0. \tag{4}$$

The resolution of the test is addressed below by considering an asymptotic approach. Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be two samples of random intervals being independent and identically distributed as $X$ and $Y$, respectively. The test statistic is defined as

$$T_n = \sqrt{n} \max \Big\{d\,\text{spr}\overline{Y_n} - \text{spr}\overline{X_n}, d\,\text{spr}\overline{X_n} - \text{spr}\overline{Y_n}, \\ (1+d)\,\left|\text{mid}\overline{X_n} - \text{mid}\overline{Y_n}\right| + (d-1)\left(\text{spr}\overline{X_n} + \text{spr}\overline{Y_n}\right)\Big\}. \tag{5}$$

From now on, let us consider the bivariate normal distributions $Z = (z_1, z_2)^T \equiv \mathcal{N}_2\Big(\mathbf{0}, \Sigma_1\Big)$ and $U = (u_1, u_2)^T \equiv \mathcal{N}_2\Big(\mathbf{0}, \Sigma_2\Big)$, where $\Sigma_1$ is the covariance matrix for the random vector $(\text{mid}X, \text{spr}X)$ and $\Sigma_2$ is the corresponding one for $(\text{mid}Y, \text{spr}Y)$. The limit distribution of the statistic $T_n$ under $H_0$ is analyzed in the following result.

**Theorem 1** *For $n \in \mathbb{N}$, let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be simple random samples from $X$ and $Y$, respectively. Let $T_n$ be defined as in (5). If $X, Y \in \mathcal{P}$, then:*

(a) *Whenever* $\text{spr}E(X) = d\,\text{spr}E(Y)$ *and* $\text{mid}E(X) - \text{mid}E(Y) = (1-d)\text{spr}E(Y)$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{du_2 - z_2, (1+d)(z_1 - u_1) + (d-1)(z_2 + u_2)\}. \qquad (6)$$

(b) *Whenever* $\text{spr}E(X) = d\,\text{spr}E(Y)$ *and* $-\text{mid}E(X) + \text{mid}E(Y) = (1-d)\text{spr}E(Y)$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{du_2 - z_2, (1+d)(u_1 - z_1) + (d-1)(z_2 + u_2)\}. \qquad (7)$$

(c) *Whenever* $d\,\text{spr}E(X) = \text{spr}E(Y)$ *and* $\text{mid}E(X) - \text{mid}E(Y) = \frac{(1-d)}{d}\text{spr}E(Y)$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{dz_2 - u_2, (1+d)(z_1 - u_1) + (d-1)(z_2 + u_2)\}. \qquad (8)$$

(d) *Whenever* $d\,\text{spr}E(X) = \text{spr}E(Y)$ *and* $-\text{mid}E(X) + \text{mid}E(Y) = \frac{(1-d)}{d}\text{spr}E(Y)$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{dz_2 - u_2, (1+d)(u_1 - z_1) + (d-1)(z_2 + u_2)\}. \qquad (9)$$

*Proof* The statistic $T_n$ can be equivalently expressed as $T_n = \sqrt{n}\max\{A, B, C\}$, where $A = d\left(\text{spr}\overline{Y_n} - \text{spr}E(Y)\right) + d\,\text{spr}E(Y) - \text{spr}E(X) + \text{spr}E(X) - \text{spr}\overline{X_n}$, $B = d\left(\text{spr}\overline{X_n} - \text{spr}E(X)\right) + d\,\text{spr}E(X) - \text{spr}E(Y) + \text{spr}E(Y) - \text{spr}\overline{Y_n}$ and $C = (1+d)\left|\text{mid}\overline{X_n} - \text{mid}E(X) + \text{mid}E(X) - \text{mid}E(Y) + \text{mid}E(Y) - \text{mid}\overline{Y_n}\right| + (d-1)\left(\text{spr}\overline{X_n} - \text{spr}E(X) + \text{spr}E(X) + \text{spr}E(Y) - \text{spr}E(Y) + \text{spr}\overline{Y_n}\right)$

(a) If $\text{spr}E(X) = d\,\text{spr}E(Y)$ and $\text{mid}E(X) - \text{mid}E(Y) = (1-d)\text{spr}E(Y)$, the second term and the negative form of the third term diverge in probability to $-\infty$ as $n \to \infty$ by the Central Limit and the Slutsky's theorems. Finally, the Continuous Mapping and the Central Limit Theorems for real variables lead to (6).
Similar reasonings can be taken into account in the other three situations:

(b) If $\text{spr}E(X) = d\,\text{spr}E(Y)$ and $-\text{mid}E(X) + \text{mid}E(Y) = (1-d)\text{spr}E(Y)$, the second term and the positive form of the third term diverges in probability to $-\infty$ as $n \to \infty$;

(c) If $d\,\text{spr}E(X) = \text{spr}E(Y)$ and $\text{mid}E(X) - \text{mid}E(Y) = \frac{(1-d)}{d}\text{spr}E(Y)$, the first term and the negative form of the third term diverges in probability to $-\infty$ as $n \to \infty$;

(d) If $d\,\text{spr}E(X) = \text{spr}E(Y)$ and $-\text{mid}E(X) + \text{mid}E(Y) = \frac{(1-d)}{d}\text{spr}E(Y)$, the first term and the negative form of the third term diverges in probability to $-\infty$ as $n \to \infty$.

□

*Remark 1* As in the real framework, other situations under $H_0$ being different than the ones shown in Theorem 1 (which are the 'worst' or 'limit' situations under $H_0$) lead the statistic $T_n$ to converge weakly to a limit distribution which is stochastically bounded for one of those provided in the theorem.

Since the limit distribution of $T_n$ depends on $X$ and $Y$, we can consider the following $(X, Y)$-dependent distribution for the theoretical analysis of the testing procedure (see [13]):

$$\begin{aligned} T_n' = \max\Big\{ & \sqrt{n}\left(d\ \left(\text{spr}\overline{Y_n} - \text{spr}E(Y)\right) + \text{spr}E(X) - \text{spr}\overline{X_n}\right) \\ & + \min\left(0, n^{1/4}(\text{spr}\overline{Y_n} - \text{spr}\overline{X_n})\right), \\ & \sqrt{n}\left(d\ \left(\text{spr}\overline{X_n} - \text{spr}E(X)\right) + \text{spr}E(Y) - \text{spr}\overline{Y_n}\right) \\ & + \min\left(0, n^{1/4}(\text{spr}\overline{X_n} - \text{spr}\overline{Y_n})\right), \\ & \sqrt{n}\Big((1+d)\left(\text{mid}\overline{X_n} - \text{mid}E(X) + \text{mid}E(Y) - \text{mid}\overline{Y_n}\right) \\ & \quad + (d-1)\left(\text{spr}\overline{X_n} - \text{spr}E(X) + \text{spr}\overline{Y_n} - \text{spr}E(Y)\right)\Big) \\ & + \min\left(0, n^{1/4}(\text{mid}\overline{X_n} - \text{mid}\overline{Y_n})\right), \\ & \sqrt{n}\Big((1+d)\left(\text{mid}E(X) - \text{mid}\overline{X_n} - \text{mid}E(Y) + \text{mid}\overline{Y_n}\right) \\ & \quad + (d-1)\left(\text{spr}\overline{X_n} - \text{spr}E(X) + \text{spr}\overline{Y_n} - \text{spr}E(Y)\right)\Big) \\ & + \min\left(0, n^{1/4}(\text{mid}\overline{Y_n} - \text{mid}\overline{X_n})\right)\Big\}. \end{aligned} \tag{10}$$

As in [13], the inclusion of $\min\left(0, n^{1/4}(\text{spr}\overline{Y_n} - \text{spr}\overline{X_n})\right)$ (and so for *mids*) in $T_n'$ are useful to determine the terms on its expression having relevance depending on each situation considered under $H_0$. The consistency and the power of the test are shown in Theorem 2.

**Theorem 2** *Let $\alpha \in [0, 1]$ and $k_{1-\alpha}$ be the $(1-\alpha)$-quantile of the asymptotic distribution of $T_n'$. If $H_0$ in (4) is true, then it is satisfied that*

$$\limsup_{n\to\infty} P\left(T_n' > k_{1-\alpha}\right) \leq \alpha,$$

*and the equality is achieved whenever conditions in a), b), c) and d) in Theorem 1 are fulfilled. In addition, if $H_0$ is not true, then*

$$\lim_{n\to\infty} P\left(T_n' > k_{1-\alpha}\right) = 1.$$

As an immediate consequence of Theorem 2, the test which rejects $H_0$ in (4) at the significance level $\alpha$ whenever $T_n' > k_{1-\alpha}$ is asymptotically efficient and consistent.

### 3.1 Bootstrap Test

Since the asymptotic limit distribution is not easy to handle in practice, a residual bootstrap approach is proposed. Let $X$ and $Y$ be two RIs such that $\text{spr}E(X) > 0$ and $\text{spr}E(Y) > 0$, and let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be two simple random samples drawn from $X$ and $Y$, respectively. Let us consider bootstrap samples for $X$ and $Y$, i.e. $\{X_i^*\}_{i=1}^n$ and $\{Y_i^*\}_{i=1}^n$ being chosen randomly and with replacement from $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively. The bootstrap statistic is based on the expression of $T_n'$ and it is defined as follows:

$$
\begin{aligned}
T_n^* = \max \Big\{ & \sqrt{n}\left(d\ \left(\text{spr}\overline{Y_n^*} - \text{spr}\overline{Y_n}\right) + \text{spr}\overline{X_n} - \text{spr}\overline{X_n^*}\right) \\
& + \min\left(0, n^{1/4}(\text{spr}\overline{Y_n} - \text{spr}\overline{X_n})\right), \\
& \sqrt{n}\left(d\ \left(\text{spr}\overline{X_n^*} - \text{spr}\overline{X_n}\right) + \text{spr}\overline{Y_n} - \text{spr}\overline{Y_n^*}\right) \\
& + \min\left(0, n^{1/4}(\text{spr}\overline{X_n} - \text{spr}\overline{Y_n})\right), \\
& \sqrt{n}\Big((1+d)\left(\text{mid}\overline{X_n^*} - \text{mid}\overline{X_n} + \text{mid}\overline{Y_n} - \text{mid}\overline{Y_n^*}\right) \\
& \qquad + (d-1)\left(\text{spr}\overline{X_n^*} - \text{spr}\overline{X_n} + \text{spr}\overline{Y_n^*} - \text{spr}\overline{Y_n}\right)\Big) \\
& + \min\left(0, n^{1/4}(\text{mid}\overline{X_n} - \text{mid}\overline{Y_n})\right), \\
& \sqrt{n}\Big((1+d)\left(\text{mid}\overline{X_n} - \text{mid}\overline{X_n^*} + \text{mid}\overline{Y_n^*} - \text{mid}\overline{Y_n}\right) \\
& \qquad + (d-1)\left(\text{spr}\overline{X_n^*} - \text{spr}\overline{X_n} + \text{mid}\overline{Y_n^*} - \text{mid}\overline{Y_n}\right)\Big) \\
& + \min\left(0, n^{1/4}(\text{mid}\overline{Y_n} - \text{mid}\overline{X_n})\right) \Big\}.
\end{aligned} \tag{11}
$$

The different asymptotic distributions of $T_n^*$ are (almost sure) the ones provided in Theorem 1 for $T_n$, under the same conditions, and the consistency of the bootstrap procedure is straightforwardly derived. The distribution of $T_n^*$ is approximated in practice by means of the Monte Carlo method.

## 4 Simulations

The empirical behaviour of the bootstrap test is shown by simulation. Two different situations are considered: in the first one the *mid* and *spr* components of the two independent RIs $X$ and $Y$ are independently generated. In the second situation, it is allowed that those components have certain level of dependence each other. The two situations are described as follows:

- **Situation 1**: $\text{mid}X \equiv \mathcal{N}(2, 5)$, $\text{spr}X \equiv U(1, 3)$; $\text{mid}Y \equiv \mathcal{N}(3, 5)$, $\text{spr}Y \equiv U(1, 5)$.
- **Situation 2**: $\text{mid}X \equiv U(2, 6)$, $\text{spr}X \equiv \text{mid}X/2$; $\text{mid}Y = \text{spr}Y \equiv U(1, 5)$.

It is straightforward to show that the theoretical situation 1 satisfies the conditions (a) of Theorem 1, and the situation 2 is under conditions (b). Besides, $S(E(X), E(Y)) = 2/3$ in both cases.

**Table 1** Empirical size of the two-sample similarity bootstrap test in Situations 1 and 2

| $n\diagdown 100\cdot\alpha$ | Situation 1 | | | Situation 2 | | |
|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 |
| 10 | 2.27 | 6.88 | 10.64 | 2.38 | 7.36 | 11.85 |
| 30 | 1.60 | 4.60 | 9.54 | 1.89 | 5.73 | 10.31 |
| 50 | 1.35 | 4.96 | 10.59 | 1.44 | 5.32 | 10.46 |
| 100 | 1.27 | 5.06 | 10.35 | 1.22 | 5.18 | 10.24 |
| 200 | 0.95 | 4.89 | 9.88 | 1.1 | 5.12 | 9.8 |

The bootstrap test proposed in Sect. 3.1 has been run for 10000 simulations with 1000 bootstrap replications each to test $H_0 : S(E(X), E(Y)) \geq 2/3$ versus $H_1 : S(E(X), E(Y)) < 2/3$, for several significance levels $\alpha$ and different sample sizes. Results are gathered in Table 1. They show that the empirical sizes of the test are in both cases quite close to the expected nominal significance levels even for moderate sample sizes. Specifically, the approximation to the nominal significance level is more conservative in the first situation than in the second one. The slight differences appreciated in the two situations may be due to the diverse nature of the distributions.

Finally, a small empirical study to show the power of the proposed test has been developed. Specifically, mid$X$ in Case 1 has been chosen to have distributions $N(1, 5)$, $N(0, 5)$ and $N(-1, 5)$, respectively. In these cases, the bootstrap approach for $\alpha = 0.05$ and $n = 50$ lead to $p$-values of 0.153, 0.381 and 0.692, respectively, and, therefore, in this case the power of the test approximate to 1 as the distribution of $X$ moves further away from the null hypothesis.

## 5 Conclusions and Open Problems

A hypothesis test for checking the similarity between the expected value of two RIs has been introduced. A test statistic has been proposed and its limit distribution has been analyzed by means of both asymptotic and bootstrap techniques. Some simulation studies have been carried out to show the suitability of the bootstrap approach for moderate/large sample sizes.

As future work, theoretical and empirical comparisons between different similarity indexes should be developed. The power of the proposed test may also be theoretically analyzed as well as the sensitivity of the test when different distributions are chosen. Other versions of the test statistic involving the covariance matrix can be studied. Finally, the proposed test could be extended to more than two RIs and to the fuzzy framework.

**Acknowledgments** The research in this paper is partially supported by the Spanish National Grant MTM2013-44212-P, and the Regional Grant FC-15-GRUPIN-14-005. Their financial support is gratefully acknowledged.

## References

1. Aumann RJ (1965) Integrals of set-valued functions. J Math Anal Appl 12:1–12
2. Blanco-Fernndez A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput Stat Data Anal 55(9):2568–2578
3. Ferraro MB, Coppi R, González-Rodríguez G, Colubi A (2010) A linear regression model for imprecise response. Int J Approximate Reasoning 51(7):759–770
4. González-Rodríguez G, Colubi A, Gil MA (2012) Fuzzy data treated as functional data: a one-way ANOVA test approach. Comput Stat Data Anal 56(4):943–955
5. Horowitz JL, Manski CF (2006) Identification and estimation of statistical functionals using incomplete data. J Econ 132:445–459
6. Hudgens MG (2005) On nonparametric maximum likelihood estimation with interval censoring and left truncation. J Roy Stat Soc: Ser B 67:573–587
7. Jaccard P (1901) tude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Socit Vaudoise des Sciences Naturelles 37:547–579
8. Körner R (2000) An asymptotic $\alpha$-test for the expectation of random fuzzy variables. J Stat Plann Infer 83:331–346
9. Magnac T, Maurin E (2008) Partial identification in monotone binary models: discrete regressors and interval data. Rev Econ Stud 75:835–864
10. Matheron G (1975) Random sets and integral geometry. Wiley, New York
11. Molchanov I (2005) Theory of random sets. Springer, London
12. Ramos-Guajardo AB, Colubi A, González-Rodríguez G (2014) Inclusion degree tests for the Aumann expectation of a random interval. Inf Sci 288(20):412–422
13. Ramos-Guajardo AB (2015) Similarity test for the expectation of a random interval and a fixed interval. In: Grzegorzewski P, Gagolewski M, Hryniewicz O, Gil MA (eds) Strengthening links between data analysis and soft computing. Adv Intell Syst Comput 315:175–182
14. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

# Handling Uncertainty in Structural Equation Modeling

**Rosaria Romano and Francesco Palumbo**

**Abstract** This paper attempts to propose an overview of a recent method named *partial possibilistic regression path modeling* (PPRPM), which is a particular *structural equation model* that combines the principles of *path modeling* with those of *possibilistic regression* to model the net of relations among variables. PPRPM assumes that the *randomness* can be referred to the *measurement error*, that is the error in modeling the relations among the observed variables, and the *vagueness* to the *structural error*, that is the uncertainty in modeling the relations among the latent variables behind each block. PPRPM gives rise to possibilistic regressions that account for the imprecise nature or vagueness in our understanding phenomena, which is manifested by yielding interval path coefficients of the structural model. However, possibilistic regression is known to be a model sensitive to extreme values. That is way recent developments of PPRPM are focused on robust procedures for the detection of extreme values to omit or lessen their effect on the modeling. A case study on the motivational and emotional aspects of teaching is used to illustrate the procedure.

**Keywords** Interval-valued data · Possibilistic regression · SEM · Extreme values

## 1 Introduction

Path Analysis (PA) represents a widely used tool in exploratory and confirmatory statistical analysis to describe direct dependencies among set of variables [10]. A special class of PA is represented by Structural Equation Models (SEM) [4], which aim to estimate a network of causal relationships among latent variables (LVs) defined by blocks of manifest variables (MVs). The relations among the LVs define the *structural model*, whereas the relations between each LV and its own block of MVs

R. Romano (✉)
University of Calabria, Cosenza, Italy
e-mail: rosaria.romano@unical.it

F. Palumbo
University of Naples Federico II, Naples, Italy
e-mail: fpalumbo@unina.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_53

define the *measurement model*. The common features of PA and SEM are that: (i) two or more sets of variables are involved; (ii) at least one of these variables is latent. SEM are generally divided into two categories, according to the estimation procedure [21]: covariance based SEM (CBSEM) and variance based SEM (VBSEM). CBSEM estimates the model parameters through a unique estimation of the Variance-Covariance matrix. Under the usual assumptions, estimation is achieved via Maximum Likelihood (ML) approach. VBSEM estimation is a two-step procedure. Partial Least Squares Path Modeling (PLSPM) is the most largely used approach for VBSEM that partially estimates the outer model parameters and the inner model parameters alternatively [22]. Each block is estimated independently and the procedure stops when the convergence is reached. Albeit the original proposal is based on ordinary least squares estimation for both the outer and the inner model parameters, several different approaches have been proposed. In particular some alternatives are based on the least absolute deviation (LAD) regression. For sake of space, we skip any discussion on the model interpretation in CBSEM and VBSEM and we focus on the model residuals.

The model goodness of fit in CBSEM is evaluated by comparing the observed Variance-Covariance matrix and the estimated one. In VBSEM the attention is focused on the residual, defined as the deviation between the estimated and the observed dependent variable. In such a case, we deal with two different sources of residuals: the outer model and the inner model residuals. Outer model residuals are interpretable using the usual reading-key, but the inner model residual cannot. Inner model residuals represent the model inadequacy in describing the relationships between the latent variables. Recalling George Box: "*All models are wrong but some are useful*", so the model inadequacy can be described by the vagueness of its parameters. The possibilistic theory approach allows us to take into account the vagueness by interval-valued parameters.

Recently, a new method named Partial Possibilistic Regression Path Modeling (PPRPM) [14–16] has been proposed as an alternative to the classical PLSPM. As discussed in [15], PPRPM is aiming at dealing with the two sources of uncertainty in the VBSEM: (a) the *measurement error* related to the relations between each LV and its own block of items, (b) the *structural error* related to the relations among the LVs. The former is generally defined as any deviation from the true value of a variable that arises in the measurement process. The latter is something different: it originates from the relationships between variables that are latent and not directly measured. PPRPM assumes that the *randomness* can be referred to the *measurement error* and the *vagueness* to the *structural error*. The main idea is that variability in structural model is not caused by the error but by the intrinsic variety of the systems output.

PPRPM differently minimizes the two error components. The *randomness* is minimized in the same way as the classical PLSPM approach based on classical linear regressions, but using the least absolute values instead of the least squares. The *vagueness* is minimized by the Possibilistic Regression (PR) [19], which considers this type of uncertainty as included in the range of model parameters, defined as interval-valued data [1], i.e. range of values denoted in terms of midpoint and range. The estimation procedure consists in solving a problem whose objective is to minimize the range of the interval-valued parameters. This choice allows us to take into account the vague relations among the LVs, on the one hand, and on the other hand, the use of the least absolute values allows us to get a more robust estimate of the LV scores and ensures consistency between the minimization procedure of the two error components. In fact, PPRPM estimation process is an $L^1$ norm problem that independently minimizes the sum of the absolute values of the residuals in the measurement model and the sum of all the ranges of the interval-valued coefficients in the structural model.

PR was introduced by Tanaka and Watada [20], who established their idea on the basis of possibility theory [23]. Since then different approaches have been proposed to cope with vagueness in regression analysis. For the sake of simplicity they can be grouped into two broad categories: Fuzzy Least Square Regression (FLSR) and Possibilistic Regression (PR). Two papers can be considered seminal for each approach, while many others have proposed further developments. Diamond's papers [7, 8] introduced the FLSR approach (see also [5, 6]) which has been extended to the interval data analysis [2, 3, 12] and to symbolic data analysis [9]. The paper by Tanaka and Asai [18] introduced the PR approach. We refer the reader to the book by Tanaka and Guo [19] for an exhaustive overview of possibilistic data analysis.

Despite the new developments, Tanaka's approach remains the benchmark as a model for handling vagueness in case of crisp data. That is way PR is used in PPRPM to model structural relations. But it is excessively sensitive to extreme values leading to broad interval outputs that may make results inaccurate for a useful interpretation.

This proposal aims to propose an overview of the use of PPRPM. Recent developments focus on a *robustifying* procedure to PPRPM, where according to [17] extreme values are detected to omit or lessen their effect.

The paper is organized as follows: Sect. 2 shortly introduces the interval data notation, summarizes the PPRPM and illustrates a procedure to handle extreme values in PR; Sect. 3 shows an example on real data.

## 2 Partial Possibilistic Regression Path Modeling

Interval-valued data represent a special case of fuzzy data, generally defined in terms of extreme values (lower and upper bound) or midpoint and range. A rigorous study of *interval data* is given by *Interval Analysis* [1]. In this framework, an *interval value* is a bounded subset of real numbers $[x] = [\underline{x}, \overline{x}]$:

$$[x] = \{x \in \mathbf{R} | \; \underline{x} \leq x \leq \overline{x}\}, \tag{1}$$

where $\underline{x}$ and $\overline{x}$ are the *lower* and *upper bound*, respectively. Alternatively the *range/midpoint* notation is defined as:

$$r(x) = |\overline{x} - \underline{x}| \qquad\qquad c(x) = \frac{1}{2}|\underline{x} + \overline{x}|,$$

where $r(x)$ and $c(x)$ refer to the range and the midpoint, respectively. For sake of short notation, the set $\{c(x); r(x)\}$ can also be noted as $\tilde{x}$ (or $\{c, r\}$).

PR defines the relation between one dependent variable $Y$ and a set of $M$ predictors $X_1, X_2, \ldots, X_M$, observed on $N$ statistical units, through a linear function holding interval valued coefficients [20]

$$Y = \tilde{\omega}_1 X_1 + \cdots + \tilde{\omega}_m X_m + \cdots + \tilde{\omega}_M X_M, \tag{2}$$

where $\tilde{\omega}_m$ denotes the generic interval-valued coefficient in terms of midpoint and range: $\tilde{\omega}_m = \{c_m; r_m\}$. It is worth noting that there is no error term in Eq. 2, since the interval-valued coefficients $\tilde{\omega}_m$ embed it. PR aims to minimize the sum of the interval coefficient ranges

$$min \textstyle\sum_{m=1}^{M} \left(\sum_{n=1}^{N} r_m |x_{nm}|\right), \tag{3}$$

under the following linear constraints

$$\begin{aligned} &\textstyle\sum_{m=1}^{M} c_m x_{nm} + \sum_{m=1}^{M} r_m |x_{nm}| \geq y_n, \\ &\textstyle\sum_{m=1}^{M} c_m x_{nm} - \sum_{m=1}^{M} r_m |x_{nm}| \leq y_n, \qquad \forall n = 1, \ldots, N, \end{aligned} \tag{4}$$

satisfying the following conditions: (i) $r_m \geq 0$; (ii) $c_m \in R$. Constraints in (4) guarantee the inclusion of the whole given data in the estimated boundaries. In a geometric view, where statistical units are represented as points in the $\Re^{M+1}$ space, the optimal solution ensures the inclusion of the whole given data set in the estimated boundaries with the minimum range of parameters.

PPRPM estimation process is an $L^1$ norm problem that independently minimizes the *sum of the absolute values of the residuals* in the measurement model and the *sum of all the ranges of the interval-valued coefficients* in the structural model. The algorithm computes the latent variables' scores alternating the *outer* and *inner* estimation till convergence. The procedure starts on centered (or standardized) MVs by choosing arbitrary weights $w_{ph}$. In the external estimation, the LV is estimated as a linear combination of its own MVs:

$$\mathbf{v}_h \propto \sum_{p=1}^{P_h} w_{ph} \mathbf{x}_{ph} = \mathbf{X}_h \mathbf{w}_h, \tag{5}$$

where $\mathbf{v}_h$ is the standardized outer estimation of the LV $\xi_h$ and the symbol $\propto$ means that the left-hand side of the equation corresponds to the standardized right-hand side. In the internal estimation, the LV is estimated by considering its links with the other adjacent $H'$ latent variables:

$$\vartheta_h \propto \sum_{h'=1}^{H'} e_{hh'}\mathbf{v}_{h'}, \tag{6}$$

where $\vartheta_h$ is the standardized inner estimation of the LV $\xi_h$ and the inner weights, according to the so called *centroid scheme* [11], are equal to the sign of the correlation between the outer estimate $\mathbf{v}_h$ of the $h$th LV and the outer estimate of the $h'$ LV $\mathbf{v}_{h'}$ connected with $\mathbf{v}_h$.

These first two steps allow us to update the outer weights $w_{ph}$. The weight $w_{ph}$ is the regression coefficient in the median regression of the $p$th MV of the $h$th block $\mathbf{x}_{ph}$ on the inner estimate of the $h$th LV $\vartheta_h$:

$$\mathbf{x}_{ph} = w_{ph}\vartheta_h + \epsilon_{ph}. \tag{7}$$

The algorithm iterates till convergence and it is demonstrated to be convergent for one and two-block models. However, for multi-block models, convergence is always verified in practice. After convergence, structural (or path) coefficients are estimated through PR among the estimated LVs.

$$\xi_j = \tilde{\beta}_{0j} + \sum_{h:\xi_h \to \xi_j} \tilde{\beta}_{hj}\xi_h, \tag{8}$$

where $\xi_j (j = 1, \dots, J)$ is the generic endogenous (dependent) LV and $\tilde{\beta}_{hj}$ is the generic *interval path coefficient* or equivalently $[\underline{\beta}_{hj}, \overline{\beta}_{hj}] = [c_{hj} \pm a_{hj}]$, interrelating the $h$th exogenous (independent) variable to the $j$th endogenous one. The higher the midpoint coefficient the higher the contribution to the prediction of the endogenous LV, while the higher the range coefficient the higher the imprecision in the relation among the considered LVs.

As discussed in Sect. 1, PR is sensitive to extreme values. A recent contribution [17] has shown a procedure to handle outliers in PR. The proposed approach has been implemented in PPRPM. The *robustifying* procedure begins once PPRPM have reached convergence. Each structural equation is undergone to the following steps:

1. run the OLS on all of the LV's scores;
2. if the amount of $R$-square is $\geq 0.8$ go to step 6, else go to step 3;
3. In turn, from first to end, put away one observation and fit a curve by OLS to the other remaining data, while keeping the corresponding $R$-square in each phase;
4. delete the observations by ignorance of which the maximum of $R$-square is reached;
5. implement OLS to the new data and go back to step 2;
6. substitute $f_n = \tilde{Y}(x_n)$ computed by the final OLS;
7. implement the PR.

# 3 Example

The case study investigates some dimensions that affect the quality of teaching in high school. In particular, we examined the motivational and emotional aspects of teachers depending on: (a) the type of high school; (b) their working position; (c) the gender; (d) the socio-cultural context in which the teacher operates. The MESI (Motivation, Emotions, Strategies, Teaching) questionnaire was used in [13], which consists of six psychometrics scales that investigate job satisfaction, practices, teaching strategies, emotions, self-efficacy, and incrementality. According to theoretical assumptions, we propose an empirical framework (see Fig. 1) for analyzing the relationships among four out of six scales composing the MESI. In our simplified MESI model, the attention is focused on the relations between satisfaction and emotions, and satisfaction and self-efficacy. Results are shown in Table 1. As can be seen, there is no relation between satisfaction and self-efficacy, since the path coefficient is equal to 0. Teach-emotions is positively related to satisfaction with a path coefficient equal to 0.69, which means that when a teacher is satisfied he/she feels more frequently positive emotions while teaching. Both satisfaction and teach-emotions are good predictors of role-emotions, with path coefficients equal to 0.39 and 0.22, respectively. In other words, when a teacher is satisfied he/she feels more frequently positive emotions also in his/her role as a teacher. In addition, the increase in positive

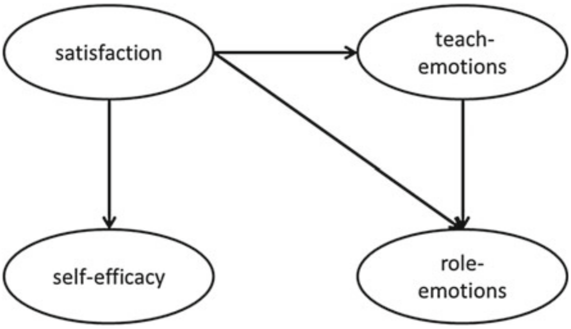


**Fig. 1** Structural model of the MESI questionnaire: the model considers the relations between satisfaction and emotions, and satisfaction and self-efficacy

**Table 1** PLSPM and PPRPM structural model results

| Relations | PLSPM path | $R^2$ | PPRPM path | IC |
|---|---|---|---|---|
| Satisfaction > self-efficacy | 0.21 | 0.05 | {0.0; 0.0} | 0.77 |
| Satisfaction > teach-emotions | 0.60 | 0.37 | {0.69; 0.23} | 0.88 |
| Satisfaction > role-emotions | 0.27 | 0.59 | {0.39; 0.0} | 0.80 |
| Teach-emotions > role emotions | 0.56 | | {0.22; 0.16} | |

emotions while teaching also increases positive emotions in the role of teacher. It is worth noting that some relations indicate a certain imprecision. This holds for the relationship between satisfaction and teach-emotions, whose path coefficient has a range equal to 0.23, and the relationship between the latter and the role-emotion, whose path coefficient has a range equal to 0.16. In Table 1 the results of the PPRPM are also compared with those of the classical PLSPM. In particular, the table shows the values of the path coefficients and of the goodness of fit indexes. As can be seen, PPRPM results are consistent with the results obtained on the classical single valued parameters model. The weak relationship between satisfaction and self-efficacy highlighted by a path coefficient close to zero in the PPRPM approach, is underlined by the low value of the $R^2$ index in PLSPM. The coefficient between satisfaction and teach-emotions is very similar in the two approaches, but PPRPM also provides information on the uncertainty of the relation. In other words, the range of the coefficient shows that the variation in the opinions of the respondents with respect to these two scales is not sufficient to arrive at a precise measurement of the dependent relationship between the two scales. Finally, both approaches show that role-emotions depend on the satisfaction and teach-emotions, but the PPRPM approach highlights the fact that there is a greater margin of imprecision in the second relation (higher range).

## 4 Concluding Remarks

This paper has shown how the proposed procedure can be considered a valid alternative to the classical SEM for analyzing ordinal subjective data. In this paper, PPRPM permits to appreciate how much the inner model (structural model) relationships are vague. However, it is well known that models based on PR are sensitive to outliers. In such a context, the present proposal has implemented the procedure proposed by [17]. For sake of space we did not discuss any detail about the procedure implementation. It requires subjective choices of the thresholds for the detection of the outliers. Current research are focused on alternative approaches to cope with such a issue.

## References

1. Alefeld G, Mayer G (2000) Interval analysis: theory and applications. J Comput Appl Math 121:421–464
2. Billard L, Diday E (2000) Regression analysis for interval-valued data. In: Data analysis, classification, and related methods. Springer, Berlin, pp 369–374
3. Blanco-Fernndez A, Corral N, Gonzguez G, (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput Stat Data Anal 55(9):2568–2578
4. Bollen K (1989) Structural equations with latent variables. Wiley, New York
5. Coppi R, D'Urso P, Giordani P, Santoro A (2006) Least squares estimation of a linear regression model with LR fuzzy. Comput Stat Data Anal 51:267–286

6. Coppi R (2008) Management of uncertainty in statistical reasoning: the case of regression analysis. Int J Approximate Reasoning 47:284–305
7. Diamond P (1988) Fuzzy least squares. Inf Sci 46:141–157
8. Diamond P (1990) Least squares fitting of compact set-valued data. J Math Anal Appl 147:531–544
9. Lima Neto EA, de Carvalho FAT (2010) Constrained linear regression models for symbolic interval-valued variables. Comput Stat Data Anal 54:333–347
10. Loehlin JC (2004) Latent variable models: an introduction to factor, path, and structural equation analysis. Erlbaum, Hillside
11. Löhmoller J (1989) Latent variable path modeling with partial least squares. Physica-Verlag, Heildelberg
12. Marino M, Palumbo F (2002) Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression. Statistica Applicata (Ital J Appl Stat) 14:277–291
13. Moé A, Pazzaglia F, Friso G (2010) MESI. Motivazioni, emozioni, strategie e insegnamento. Questionari metacognitivi per insegnanti. Edizioni Erickson
14. Romano R, Palumbo F (2013) Partial possibilistic regression path modeling for subjective measurement. J Methodol Appl Stat 15:177–190
15. Romano R, Palumbo F (2016) Partial possibilistic regression path modeling. In: Abdi et al. (eds) The multiple facets of partial least squares methods, Springer proceedings in mathematics and statistics. Springer, USA
16. Romano R, Palumbo F (2016) Comparing partial least squares and partial possibilistic regression path modeling to likert type scales: results from a montecarlo simulation study. IFCS 2016 Proceedings
17. Shakouri G, Nadimi R (2013) Outlier detection in fuzzy linear regression with crisp input-output by linguistic variable view. Appl Soft Comput 13(1):734–742
18. Tanaka H, Asai K (1982) Linear regression analysis with fuzzy model. IEEE Trans Syst Man Cybern 12:903–907
19. Tanaka H, Guo P (1999) Possibilistic data analysis for operations research. Physica-Verlag, Wurzburg
20. Tanaka H, Watada J (1987) Possibilistic linear systems and their application to the linear regression model. Fuzzy Sets Syst 27:275–289
21. Vilares MJ, Almeida MH, Coelho PS (2010) Comparison of likelihood and PLS estimators for structural equation modeling: a simulation with customer satisfaction data. Handbook of partial least squares: concepts, methods and applications. Springer, Berlin, pp 289–305
22. Wold H (1966) Estimation of principal component and related models by iterative least squares. In: Krishnaiah P (ed) Analysis multivariate. Academic Press, New York, pp 391–420
23. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1(1):3–28

# Detecting Inconsistencies in Revision Problems

**Fabian Schmidt, Jörg Gebhardt and Rudolf Kruse**

**Abstract** When dealing with complex knowledge, inconsistencies become a big problem. One important aspect of handling inconsistencies is their detection. In this paper we consider approaches to detect different types of inconsistencies that may occur in the formulation of revision problems. The general discussion focuses on the revision of probability distributions. In our practical analysis, we refer to probability distributions represented as Markov networks.

## 1 Introduction

One important aspect of maintaining knowledge for knowledge based systems is the ability to react to changes in beliefs quickly and frequently. Therefore, methods have been developed to properly adapt knowledge to new beliefs. One important aspect of proper adaptation is formulated in the **principle of minimal change** [9], which states that in order to incorporate given new beliefs, only absolutely necessary changes have to be made in a knowledge base. This means, after the incorporation of the new beliefs, the knowledge base should be as close to the original one as possible, in an information theoretic sense. The **revision operation** has been introduced as a belief change operation that applies new beliefs respecting this principle [7]. From the perspective of knowledge based systems, further properties a revision operation should satisfy have been formulated as postulates in [1, 5, 13]. How to approach revision algorithmically has been outlined in [6], and computational considerations have been made in [18]. Our work focuses on the revision of probability distributions as it has been introduced in [10]. In this context the revision operation has been

F. Schmidt (✉) · J. Gebhardt
ISC Gebhardt, Celle, Germany
e-mail: schmidt@isc-gebhardt.de

J. Gebhardt
e-mail: gebhardt@isc-gebhardt.de

R. Kruse
Otto-von-Guericke University, Magdeburg, Germany
e-mail: kruse@iws.cs.uni-magdeburg.de

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_54

successfully implemented for Markov networks [2, 11] using iterative proportional fitting [21, 23]. This method is well known in the area of statistics and shows beneficial properties for our context. Markov networks are a suitable tool to decompose high-dimensional probability spaces into a number of smaller low-dimensional probability distributions. They belong to a group of techniques called graphical models [15, 16, 19, 24].

The growing complexity and interconnectedness of knowledge bases and an increasing number of new beliefs lead almost inevitably to inconsistencies in the formulation of revision problems. In almost any type of knowledge based systems, inconsistencies render the underlying upon useless and should consequently be addressed. In this contribution we focus on inconsistencies during the revision of probability distributions. This is a multi-facet problem and different aspects of it have been introduced in [22]. Furthermore, two types of inconsistencies and a revision control algorithm have been described in [12].

In this work we focus on the important aspect of detecting the presence of inconsistencies in a given revision problem. In Sect. 2 of this paper, we will formally introduce the revision operation, specify what a revision problem is, and define revision inconsistencies. Section 3 then discusses how the problem of detecting inconsistencies can be approached, deals with different classes of possible solutions as well as a short analysis on the usability of the given classes in our scenario. In Sect. 4 we look at the detection of inconsistencies from the point of view of an application using Markov networks. Section 5 then concludes the paper and provides some ideas for future research.

## 2 Fundamentals

In this section we will describe the revision operation, define the revision problem, and specify what inconsistencies are in that context.

### *2.1 The Revision Operation*

This work focuses on the revision of probability distributions and we therefore define it in this context.

As mentioned before, the goal of (probabilistic) revision is to compute a posterior probability distribution which satisfies given new distribution conditions, only accepting a minimal change of the quantitative interaction structures of the underlying prior distribution.

More formally, in our setting, a revision operation (see [2, 12]) operates on a joint probability distribution $P(V)$ on a set $V = \{X_1, \dots, X_n\}$ of variables with finite domains $\Omega(X_i), i = 1, \dots, n$. The purpose of the operation is to adapt $P(V)$ to new sets of beliefs. The beliefs are formulated in a so-called **revision structure** $\Sigma =$

$(\sigma_s)_{s=1}^{S}$. This structure consists of **revision assignments** $\sigma_s$, each of which represents a low dimensional (conditional) probability assignment. The pair $(P(V), \Sigma)$ is called **revision problem**.

The result of the revision, and solution to the revision problem, is a probability distribution $P_\Sigma(V)$ which

- satisfies the revision assignments (the postulated new probabilities)
- preserves the probabilistic interaction structure as far as possible.

By preserving the interaction structure we mean that, except from the modifications induced by the revision assignments in $\Sigma$, all probabilistic dependencies of $P(V)$ are to be invariant. This requirement ensures that modifications are made according to the principle of minimal change.

It can be proven (see, e.g. [2]) that in case of existence, the solution of the revision problem $(P(V), \Sigma)$ is uniquely defined. This solution can be determined using iterative proportional fitting [23, 24]. Starting with the initial probability distribution, this process adapts the initial probability distribution iteratively, one revision assignment at the time, and converges to a limit distribution that solves the revision problem, given there are no inconsistencies.

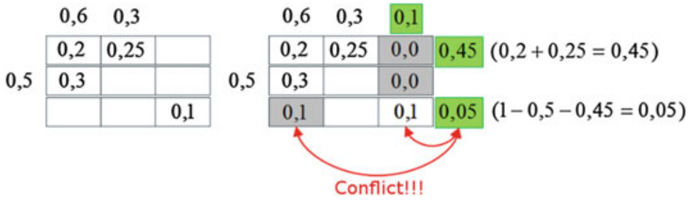
## 2.2 *Inconsistencies in the Context of the Revision Operation*

Inconsistencies in the context of revising probability distributions have been analysed in [12], and two types of inconsistencies of revision problems have been distinguished, which are *inner inconsistencies* and *outer inconsistencies*, respectively.

Inner consistency of a revision structure $\Sigma$ is given, if and only if a probability distribution exists that satisfies the revision assignments of $\Sigma$; otherwise we refer to **inner inconsistencies** of $\Sigma$.

In Fig. 1, a simple example is shown where the given revision assignments contradict each other and hence do not form a single probability distribution. The filled entries in the left table represent the revision assignments. In the right table consequences for the rest of the table are shown and one conflict is highlighted.

Given that there is a probability distribution that satisfies $\Sigma$, it is still possible that due to the zero probabilities of $P(V)$ the revision problem $(P(V), \Sigma)$ is not

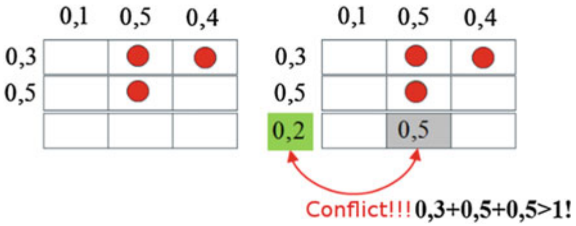


| | 0,6 | 0,3 | |
|---|---|---|---|
| | 0,2 | 0,25 | |
| 0,5 | 0,3 | | |
| | | | 0,1 |

| | 0,6 | 0,3 | 0,1 | | |
|---|---|---|---|---|---|
| | 0,2 | 0,25 | 0,0 | 0,45 | (0,2 + 0,25 = 0,45) |
| 0,5 | 0,3 | | 0,0 | | |
| | 0,1 | | 0,1 | 0,05 | (1 − 0,5 − 0,45 = 0,05) |

**Fig. 1** Inner inconsistency

**Fig. 2** Outer inconsistency



solvable. This is the case when one of those zero values would need to be modified in order to satisfy the revision assignments. Such a modification of the interaction structure of $P(V)$ is not permitted during a revision operation. Therefore, a second type of inconsistency is defined as follows:

Given that $\Sigma$ has the property of inner consistency, the revision problem $(P(V), \Sigma)$ shows the property of **outer inconsistency**, if and only if there is no solution to the revision problem.

Figure 2 illustrates an outer inconsistency. In the left table again the numbers represent revision assignments. This time there are additional circles representing zero values that cannot be changed during the revision operation. As before, the right table shows consequences for the remaining table entries as well as an inconsistency.

## 3 Detection

Detecting the presence of inconsistencies amounts to calculating the posterior probability given some evidence and is therefore NP-hard [3, 25]. Hence, to determine consistency we have to attempt the construction of a posterior probability distribution. If the construction is successful, the revision problem shows the property of consistency. This is true for both types of inconsistencies we defined earlier. In fact both problems can be transformed into one another. If one can solve the first problem, one can solve the second problem by adding revision assignments representing the zero values. The second problem is actually a generalisation of the first one - there are simply no zero values present. Hence, by solving the second problem one can solve the first one as well.

From this observation, we can infer that both problems have roughly the same degree of complexity, where the first problem most likely needs less effort to calculate. In the literature we found two general approaches to construct a high dimensional probability distribution from lower dimensional probability statements, namely algorithms that find either an *approximating solution* or *exact solutions* if there is one.

## 3.1 Approximative Algorithms

There is a whole class of algorithms for finding entropy maximising solutions based on the uniform distribution. More specifically, for Markov networks there are, for example, parameter estimation methods based on maximum likelihood and maximum entropy [8, 15, 20]. These methods are potentially faster than the exact methods and always give a result (either the exact one in case of consistency or an approximation in case of inconsistencies). In order to use this kind of methods to detect inconsistencies, one can follow a two-step process:

1. Create a candidate probability distribution
2. Check whether all revision assignments (and zero values) are satisfied

The first step is potentially faster than using an exact method. The second step, which becomes necessary because we don't know whether we have an exact solution or an approximation, may require a significant number of checks.

## 3.2 Exact Algorithms

Methods based on iterative proportional fitting that do not use approximations to speed up the process find entropy maximising solutions, can be based on any probability distribution, not just the uniform distribution. However, in case of inconsistencies there are multiple limit distributions satisfying different subsets of revision assignments. A single unique solution can only be obtained in the case of consistency. In addition to this disadvantage, they are potentially slower since they are not sacrificing accuracy for performance.

From a mathematical point of view, detecting inconsistencies with these methods is straightforward. In case of consistency the iterative proportional fitting converges towards a single unique probability distribution, which then also solves the revision problem. Otherwise, it will find multiple limit distributions, each of which is satisfying a different subset of revision assignments. In practice, the problem is to decide which of the two cases is present.

## 3.3 Further Remarks

In practical applications, detection is often embedded in the process of revising probability distributions. For that reason, it is interesting to analyse whether the constructed distributions already sufficiently solve the actual revision problem.

The approximative methods always deliver a distribution, even if inconsistencies are present. This is a useful property for working with real world problems. However, those methods maximise entropy towards the uniform distribution which is not what we need in our application. We found approaches in the literature that ,theoretically, would make those methods maximise towards a specific non-uniform distribution [4]. However, that would entail adding a large number of constraints to indicate all the deviations of the wanted prior distribution from the uniform distribution. We believe that the necessary effort then neglects the performance advantage due to the additional constraints.

The exact methods work with any kind of prior probability distribution and maximise entropy against those. If they find a unique solution, it is also a suitable solution for our revision problem. If inconsistencies are present, no unique solution can be obtained. Nevertheless, for the revision of Markov networks, an approach has been proposed in [14], that can resolve inconsistencies in a way that the resulting distribution solves the revision problem that is information theoretically closest to the original problem.

## 4 Practical Application Using Markov Networks

In our practical application we use Markov networks to efficiently represent probability distributions. In this application the detection of inconsistencies is not a separate processing step, but it is embedded in an overall revision control mechanism that detects inconsistencies, removes them and finally calculates the solution for the (then possibly) modified revision problem. Consequently, we use an exact approach based on iterative proportional fitting and the automatic elimination of inconsistencies proposed in [14].

Since we use the revision of Markov networks we can leverage the benefits of a decomposed probability distribution. This is done implicitly through the revision algorithm, which uses propagation. The propagation algorithm as described in [17] efficiently exploits the decomposition.

As mentioned previously, the problem of detecting inconsistencies in this setting is to decide whether the algorithm converges towards a single distribution or is oscillating between multiple competing distributions.

We identified several interconnected challenges when trying to decide whether convergence is reached. In industrial applications any algorithm has to deliver a result within a reasonable amount of time. Consequently, the number of iterations is usually limited. Therefore, after that limit, the algorithm has to decide whether convergence will be reached or not. We use a measure based on the sum of the differences between revision assignments and their actual value in the distribution. This method works well in many cases. However, we still have problems when the process converges slowly, or runs into a local minimum.

## 5 Conclusion

Detecting inconsistencies in revision problems is an important topic when using revision to adapt knowledge to new beliefs. In this work we discussed different approaches to detect inconsistencies in revision problems when using probabilistic revision. Both presented types of inconsistencies can be detected using very similar approaches. In this work we analysed two different classes of methods to detect inconsistencies using constructive approaches. Both classes have their advantages and disadvantages. In our setting we prefer the exact methods since, with slight modifications, they allow us to use the detection and elimination of the occurring inconsistencies in one step, and at the same time, they provide a usable solution to our revision problem. However, under different requirements approximative methods can potentially be better suited.

In the future our findings need to be verified by running tests on data from different real world applications. Furthermore, although we did not find an approach to test for inconsistencies other than to attempt the construction of a probability distribution, there might be techniques in areas like statistics that obtain a solution faster and with less calculation. Additionally, the problems with slow convergence and local minima are of interest.

## References

1. Alchourrón CE, Gärdenfors P, Makinson D (1985) On the logic of theory change: partial meet contraction and revision functions. J Symbolic Logic 50(02):510–530
2. Borgelt C, Kruse R (2004) Knowledge revision in Markov networks. Vm075031.Usc.Es, 11:93–107
3. Cooper GF (1990) The computational complexity of probabilistic inference using bayesian belief networks. Artif Intell 42(2–3):393–405 March
4. Csiszar I (1975) $I$-Divergence geometry of probability distributions and minimization problems. Ann Probab 3(1):146–158
5. Darwiche A (1997) On the logic of iterated belief revision. Artif Intell 89(1–2):1–29
6. Gabbay D (2003) Controlled revision—an algorithmic approach for belief revision. J Logic Comput 13(1):3–22
7. Gabbay D, Smets P (eds) (1998) Handbook of defeasable reasoning and uncertainty management systems, vol 3, Belief change. Kluwer Academic Press, Netherlands
8. Ganapathi V, Vickrey D, Duchi J, Koller D (2008) Constrained approximate maximum entropy learning of markov random fields. In: Proceedings of uncertainty on artificial intelligence, pp 196–203
9. Gärdenfors P (1988) Knowledge in flux: modeling the dynamics of epistemic states. MIT Press
10. Gebhardt J, Detmer H, Madsen AL (2003) Predicting parts demand in the automotive industry—an application of probabilistic graphical models. In: Proceedings of international joint conference on uncertainty in artificial intelligence
11. Gebhardt J, Klose A, Detmer H, Ruegheimer F, Kruse R (2006) Graphical models for industrial planning on complex domains. Decis Theory Multi-Agent Plann 482:131–143
12. Gebhardt J, Klose A, Wendler J (2012) Markov network revision: on the handling of inconsistencies. Computational intelligence in intelligent data analysis, vol 445 of Studies in computational intelligence. Springer, Berlin, pp 153–165

13. Katsuno H, Mendelzon AO (1991) Propositional knowledge base revision and minimal change. Artif Intell 52(3):263–294
14. Klose A, Wendler J, Gebhardt J, Detmer H (2012) Resolution of inconsistent revision problems in Markov networks. Synergies of soft computing and statistics for intelligent data analysis, vol 190 of Advances in intelligent systems and computing. Springer, Berlin, pp 517–524
15. Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press
16. Kruse R (2013) Computational intelligence., A methodological introductionSpringer, London
17. Lauritzen SL (1992) Propagation of probabilities, means and variances in mixed graphical association models. J Am Stat Assoc 87(420):1098–1108
18. Nebel B (1994) Base revision operations and schemes: representation, semantics and complexity. In: Proceedings of the eleventh European conference on artificial intelligence (ECAI94), Amsterdam, The Netherlands. Wiley, pp 341–345
19. Pearl J (1991) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann
20. Pietra SD, Pietra VD, Lafferty J (1997) Inducing features of random fields. IEEE Trans Pattern Anal Mach Intell 19(4):380–393
21. Pukelsheim F, Simeone B (2009) On the iterative proportional fitting procedure: structure of accumulation points and L1-error analysis. Structure (05):28
22. Schmidt F, Wendler J, Gebhardt J, Kruse R (2013) Handling inconsistencies in the revision of probability distributions. In: Pan J et al. (ed) HAIS13, vol 8073 of Lecture notes in computer science. Springer, Berlin, pp 598–607
23. Teh YW, Welling M (2003) On improving the efficiency of the iterative proportional fitting procedure, pp 0–7
24. Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley
25. Yu H, Engelen R (2011) Symbolic and quantitative approaches to reasoning with uncertainty: 11th European conference, ECSQARU 2011, Belfast, UK, June 29–July 1, 2011. Proceedings, chapter importance. Springer, Berlin, pp 146–157

# Tukey's Biweight Loss Function for Fuzzy Set-Valued M-estimators of Location

**Beatriz Sinova and Stefan Van Aelst**

**Abstract** The Aumann-type mean is probably the best-known measure for the location of a random fuzzy set. Despite its numerous probabilistic and statistical properties, it inherits from the mean of a real-valued random variable the high sensitivity to outliers or data changes. Several alternatives extending the concept of median to the fuzzy setting have already been proposed in the literature. Recently, the adaptation of location M-estimators has also been tackled. The expression of fuzzy-valued location M-estimators as weighted means under mild conditions allows us to guarantee that these measures take values in the space of fuzzy sets. It has already been shown that these conditions hold for the Huber and Hampel families of loss functions. In this paper, the strong consistency and the maximum finite sample breakdown point when the Tukey biweight (or bisquare) loss function is chosen are analyzed. Finally, a real-life example will illustrate the influence of the choice of the loss function on the outputs.

**Keywords** Random fuzzy set · Robustness · Location M-estimator · Bisquare loss function · Biweight loss function

## 1 Introduction

Random fuzzy sets (fuzzy random variables in Puri and Ralescu's sense [10]) are an appropriate mathematical model to formalize numerous real-life experiments characterized by an underlying imprecision. In order to analyze them statistically, a wide

B. Sinova (✉)
Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, 33007 Oviedo, Spain
e-mail: sinovabeatriz@uniovi.es

S. Van Aelst
Department of Mathematics, KU Leuven, 3001 Leuven, Belgium
e-mail: stefan.vanaelst@kuleuven.be

B. Sinova · S. Van Aelst
Department of Applied Mathematics, Computer Science and Statistics, Ghent University,
9000 Gent, Belgium

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_55

range of methods has been proposed during the last years. Unfortunately, most of this methodology is based on the Aumann-type mean, which is a well-known location measure for random fuzzy sets that fulfills many convenient properties from both the statistical and probabilistic points of view, but it presents a high sensitivity to outliers or data changes. With the aim of providing a more robust central tendency measure, several extensions of the concept of median have already been published. However, this paper focuses on the more recent and more general M-estimation approach.

Kim and Scott [9] have studied M-estimators in the kernel density estimation context, but their theory remains valid for Hilbert-valued random elements. The space of fuzzy sets can be isometrically embedded into a convex cone of a Hilbert space, which allowed us to adapt some of their results to the fuzzy-valued case in Sinova et al. [12]. Although only the one-dimensional case (random fuzzy numbers) has been specified in [12], location M-estimators can be analogously defined for random fuzzy sets and studied as in this paper.

Sufficient conditions are provided in Sinova et al. [12] to guarantee that the adaptation of Kim and Scott's results is valid, that is, that location M-estimators belong to the convex cone of the Hilbert space. Among the loss functions satisfying such assumptions, Huber's and Hampel's loss functions were analyzed in [12] to prove the strong consistency of the corresponding M-estimators and show that the maximum finite sample breakdown point is attained. Another well-known family of loss functions, Tukey's biweight (also referred to as the bisquare function), is considered in this paper. Apart from checking that the sufficient conditions also hold for this choice, the strong consistency of the Tukey location M-estimator is established and its finite sample breakdown point is derived. Proofs are based on the same sketches included for the one-dimensional case in Sinova et al. [12].

In Sect. 2, location M-estimators for random fuzzy sets are introduced and the Representer Theorem, which expresses them as weighted means under certain sufficient conditions, is recalled. In Sect. 3, the choice of Tukey's biweight loss function is analyzed in terms of the strong consistency of the resulting estimator and its finite sample breakdown point. A real-life example in Sect. 4 illustrates the influence of the choice of the loss function on the outputs. Finally, some concluding remarks are provided in Sect. 5.

## 2 Location M-estimators for Random Fuzzy Sets

In this section, location M-estimators are adapted to summarize the central tendency of random fuzzy sets. M-estimation, firstly introduced by Huber [7], is a well-established approach that yields robust estimators. The key idea behind them is to restrict the influence of outliers by substituting the square of "errors" in methods like least squared and maximum likelihood for a (usually less rapidly increasing) loss function applied to the errors of the data. The loss function, denoted by $\rho$, is usually assumed to vanish at 0 and to be even and non-decreasing for positive values.

Let $p \in \mathbb{N}$, $\mathcal{F}_c^*(\mathbb{R}^p)$ denote the space of bounded fuzzy sets and $D$ represent a metric defined on $\mathcal{F}_c^*(\mathbb{R}^p) \times \mathcal{F}_c^*(\mathbb{R}^p)$ whose associated norm fulfills the parallelogram law (which allows the isometrical embedding of $\mathcal{F}_c^*(\mathbb{R}^p)$ into the convex cone of a Hilbert space).

**Definition 1** Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R}^p)$ an associated random fuzzy set. Moreover, let $\rho$ be a continuous loss function, and $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ a simple random sample from $\mathcal{X}$. Then, the **fuzzy M-estimator of location** is the fuzzy set-valued statistic $\widehat{\widetilde{g}^M}[(\mathcal{X}_1, \ldots, \mathcal{X}_n)]$, given, if it exists, by

$$\widehat{\widetilde{g}^M}[(\mathcal{X}_1, \ldots, \mathcal{X}_n)] = \arg \min_{\widetilde{g} \in \mathcal{F}_c^*(\mathbb{R}^p)} \frac{1}{n} \sum_{i=1}^{n} \rho(D(\mathcal{X}_i, \widetilde{g})).$$

Now, a result by Kim and Scott [9] is adapted to the fuzzy-valued case. The Representer Theorem (Theorem 1) is crucial for the particularization of Kim and Scott's theory about M-estimation for the kernel density estimation problem to random fuzzy sets. The conditions they assume to ensure the existence of M-estimates of location allow us to express the M-estimates as weighted means of the sample elements and, consequently, to assure that the M-estimates are indeed fuzzy set-valued statistics.

**Theorem 1** *Consider the metric space $(\mathcal{F}_c^*(\mathbb{R}^p), D)$. Let $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ be a simple random sample from a random fuzzy set $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R}^p)$ on a probability space $(\Omega, \mathcal{A}, P)$. Moreover, let $\rho$ be a continuous loss function which satisfies the assumptions*

- *$\rho$ is non-decreasing for positive values, $\rho(0) = 0$ and $\lim_{x \to 0} \rho(x)/x = 0$,*
- *Let $\phi(x) = \rho'(x)/x$ and $\phi(0) \equiv \lim_{x \to 0} \phi(x)$, assuming that $\phi(0)$ exists and is finite.*

*Then, the M-estimator of location exists and it can be expressed as*

$$\widehat{\widetilde{g}^M}[(\mathcal{X}_1, \ldots, \mathcal{X}_n)] = \sum_{i=1}^{n} \omega_i \cdot \mathcal{X}_i$$

*with $\omega_i \geq 0$, $\sum_{i=1}^n \omega_i = 1$. Furthermore, $\omega_i \propto \phi(D(\mathcal{X}_i, \widehat{\widetilde{g}^M}[(\mathcal{X}_1, \ldots, \mathcal{X}_n)]))$.*

In Sinova et al. [12], the well-known Huber and Hampel families of loss functions were used to compute M-estimators. Recall that the *Huber loss function* [8] is given by

$$\rho_a^H(x) = \begin{cases} x^2/2 & \text{if } |x| \leq a \\ a(|x| - a/2) & \text{otherwise,} \end{cases}$$

with $a > 0$ a tuning parameter, while the *Hampel loss function* [5] corresponds to

$$\rho_{a,b,c}(x) = \begin{cases} x^2/2 & \text{if } |x| < a \\ a(|x| - a/2) & \text{if } a \leq |x| < b \\ \dfrac{a(|x|-c)^2}{2(b-c)} + \dfrac{a(b+c-a)}{2} & \text{if } b \leq |x| < c \\ \dfrac{a(b+c-a)}{2} & \text{if } c \leq |x|, \end{cases}$$

where the nonnegative parameters $a < b < c$ allow us to control the degree of suppression of large errors. The smaller their values, the higher this degree. Note that the Huber loss function is convex and puts less emphasis on large errors compared to the squared error loss. On the other hand, Hampel's loss function is not convex and can better cope with extreme outliers, since observations far from the center ($|x| \geq c$) always contribute in the same way to the loss.

Another well-known family of loss functions is the *Tukey biweight or bisquare* [1], given by:

$$\rho_c^T(x) = \begin{cases} c^2/6 \cdot (1 - (1 - (x/c)^2)^3) & \text{if } |x| \leq c \\ c^2/6 & \text{otherwise,} \end{cases}$$

with tuning parameter $c > 0$. This loss function shares with Hampel's one that it is not convex anymore and the contribution of large errors ($|x| \geq c$) to the loss does not change anymore. Therefore, the benefit of the Tukey loss function is to combine the better performance of Hampel's loss function regarding extreme outliers with the simplicity of an expression depending on just one tuning parameter, like the Huber loss function.

It can be easily checked that the family $\rho_c^T$ of loss functions fulfills all the required conditions: they are differentiable, non-decreasing for positive values and even, they vanish at 0, $\lim_{x \to 0} \rho_c^T(x)/x = 0$, $\phi_c^T(0) \equiv \lim_{x \to 0} \phi_c^T(x)$ exists and is finite.

Therefore, all the properties derived from the Representer Theorem in Sinova et al. [12] also hold when the Tukey biweight loss function is chosen. In particular, it can be highlighted that Tukey M-estimators of location are translation equivariant, but not scale equivariant in general. With the aim of avoiding the excessive influence of the measurement units on the outputs, due to the lack of scale equivariance unless $\rho$ is a power function, the tuning parameters will be selected based on the distribution of the distances to the center. That is, we first compute an initial robust estimator of location (e.g., the impartial trimmed mean as in Colubi and González-Rodríguez [2] or, if $p = 1$, the 1-norm median in Sinova et al. [11]) and then, the distances between each observation and this initial estimate are calculated. Our recommendation is to use the 1-norm median as initial estimate when analyzing random fuzzy numbers, since its computation is not complex and this measure does not depend on the existence or not of outliers in the sample to provide us with a good initial estimate. The impartial trimmed mean (see Colubi and González-Rodríguez [2]) presents the disadvantage of requiring to fix the trimming proportion "a priori" and, in case there are no outliers, the initial estimate could be a bit far from the real center of the sample distribution. The choice for the tuning parameters $a$, $b$ and $c$ will be, along this paper, the median, the 75th and the 85th percentiles of those distances, following Kim and Scott's suggestion [9].

Regarding the practical computation of Tukey M-estimators of location, recall that the standard iteratively re-weighted least squares algorithm (see, for example, Huber [7]) can provide us with an approximation as in [12]:

Step 1 Select initial weights $\omega_i^{(0)} \in \mathbb{R}$, for $i \in \{1, \ldots, n\}$, such that $\omega_i^{(0)} \geq 0$ and $\sum_{i=1}^n \omega_i^{(0)} = 1$ (which is equivalent to choose a robust estimator of location to initialize the algorithm).

Step 2 Generate a sequence $\{\widetilde{g}_{(k)}^M\}_{k\in\mathbb{N}}$ by iterating the following procedure:

$$\widetilde{g}_{(k)}^M = \sum_{i=1}^n \omega_i^{(k-1)} \mathcal{X}_i, \quad \omega_i^{(k)} = \frac{\phi_c^T(D(\mathcal{X}_i, \widetilde{g}_{(k)}^M))}{\sum_{j=1}^n \phi_c^T(D(\mathcal{X}_j, \widetilde{g}_{(k)}^M))}.$$

Step 3 Terminate the algorithm when

$$\frac{|\frac{1}{n}\sum_{i=1}^n \rho_c^T(D(\mathcal{X}_i, \widetilde{g}_{(k+1)}^M)) - \frac{1}{n}\sum_{i=1}^n \rho_c^T(D(\mathcal{X}_i, \widetilde{g}_{(k)}^M))|}{\frac{1}{n}\sum_{i=1}^n \rho_c^T(D(\mathcal{X}_i, \widetilde{g}_{(k)}^M))} < \varepsilon,$$

for some desired tolerance $\varepsilon > 0$.

## 3 Specific Properties of Fuzzy-Valued Location M-estimators Based on Tukey Biweight Loss Function

The strong consistency of fuzzy number-valued M-estimators of location was studied in Sinova et al. [12] for specific loss functions: $\rho$ being either non-decreasing for positive values, subadditive and unbounded or the Huber or Hampel loss function (independently of the values of the tuning parameters). However, this result can be generalized to cover any bounded loss function and, in consequence, the Tukey biweight choice.

**Theorem 2** *Consider the metric space $(\mathcal{F}_c(A), D)$, with $A$ a non-empty compact convex set of $\mathbb{R}^p$ and $D$ topologically equivalent to the mid/spr-based $L^2$ distance $D_\theta^\ell$ (see Trutschnig et al. [13] for details concerning this metric). Let $\mathcal{X} : \Omega \to \mathcal{F}_c(A)$ be a random fuzzy set associated with a probability space $(\Omega, \mathcal{A}, P)$. Under any of the following assumptions:*

- *$\rho$ is non-decreasing for positive values, subadditive and unbounded,*
- *$\rho$ , for positive values, has linear upper and lower bounds with the same slope,*
- *$\rho$ is bounded,*

*and whenever the associated M-location value*

$$\tilde{g}^M(\mathcal{X}) = \arg \min_{\widetilde{U} \in \mathcal{F}_c(A)} E\left[\rho\left(D(\mathcal{X}, \widetilde{U})\right)\right]$$

*exists and is unique, the M-estimator of location is a strongly consistent estimator of $\tilde{g}^M(\mathcal{X})$, i.e.,*

$$\lim_{n\to\infty} D(\widehat{\widetilde{g}^M}[(\mathcal{X}_1, \ldots, \mathcal{X}_n)], \tilde{g}^M(\mathcal{X})) = 0 \quad \textit{a.s. [P].}$$

It should be clarified that it is very common in practice to fix a bounded referential, as is the case for the fuzzy rating scale (see Hesketh et al. [6]) when $p = 1$.

With respect to the robustness of the location M-estimators based on the Tukey biweight loss function, their *finite sample breakdown point*, for short fsbp (Donoho and Huber [3], Hampel [4]) has been computed. The fsbp represents the smallest fraction of sample observations that needs to be perturbed to make the distances between the original and the contaminated M-estimates arbitrarily large.

**Theorem 3** *Consider the metric space $(\mathcal{F}_c^*(\mathbb{R}^p), D)$. Let $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R}^p)$ be a random fuzzy set associated with a probability space $(\Omega, \mathcal{A}, P)$ and let $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ be a sample obtained from $\mathcal{X}$. Moreover, let $\rho$ be a continuous loss function fulfilling the assumptions in Theorem 1, upper bounded by certain $C < \infty$ and satisfying*

$$\rho\left(\max_{1\le i,j\le n} D(\widetilde{x}_i, \widetilde{x}_j)\right) < \frac{n - 2\lfloor\frac{n-1}{2}\rfloor}{n - \lfloor\frac{n-1}{2}\rfloor - 1} \cdot C,$$

*and such that the corresponding sample M-estimate of location is unique. Then the finite sample breakdown point of the corresponding location M-estimator is exactly $\frac{1}{n}\lfloor\frac{n+1}{2}\rfloor$, where $\lfloor\cdot\rfloor$ denotes the floor function.*

## 4 Real-Life Example

A real-life example now illustrates fuzzy-valued location M-estimators.

*Example* 68 fourth grade students from Colegio San Ignacio (Oviedo, Spain) have been asked to answer some questions from the joint Student questionnaire TIMSS (Trends in International Mathematics and Science Study)—PIRLS (Progress in International Reading Literacy Study) survey using a fuzzy rating scale (Hesketh et al. [6]). To simplify the instructions given to the nine-and-ten-year-old students, only trapezoidal fuzzy numbers have been considered. This study is going to be limited to the item that represents the degree of agreement with the statement "studying mathematics is harder than any other subject".

Location M-estimators based on Huber, Hampel and Tukey loss functions have been computed using the mid/spr-based $L^2$ distance $D^\ell_{\theta=1/3}$, where $\ell$ denotes the

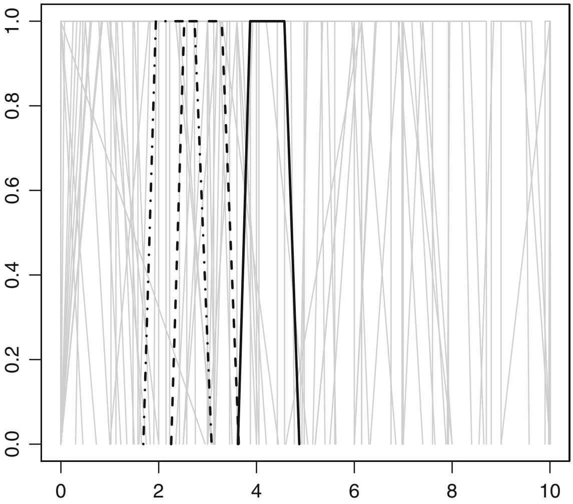


**Fig. 1** In *black*, Huber (*solid line*), Hampel (*dashed line*) and Tukey (*dash-dot line*) M-estimates for the fuzzy-valued data (in *grey*) from Example

Lebesgue measure on [0, 1] (see Trutschnig et al. [13]). The 1-norm median in [11] has been considered as the initial robust estimator for the selection of the tuning parameters and the initialization of the algorithm to approximate the M-estimates.

The outputs for the three M-estimates have been displayed in Fig. 1.

As shown in Sinova et al. [12], when analyzing trapezoidal fuzzy numbers, any loss function fulfilling the conditions stated for the Representer Theorem provides us with an M-estimate of trapezoidal shape too.

Notice that the aim of this example is just to illustrate the computation of fuzzy-valued M-estimators and the influence the choice of the loss function has on the outputs, but not to provide a comparison of the different loss functions. On one hand, there are no outliers in the answers given by the students and, on the other hand, the best choice of $\rho$ also depends on different factors (e.g., the weight we wish to assign to the outliers in each specific example or the selection of the tuning parameters).

## 5 Concluding Remarks

The Tukey biweight or bisquare family of loss functions has been used in order to compute fuzzy set-valued M-estimators of location through the Representer Theorem. The strong consistency and the robustness of this choice have been given. In future research, it would be interesting to develop a sensitivity analysis on how the selection of the involved tuning parameters affect the computation of M-estimators, as well as a deeper study of other families of loss functions for which the Representer Theorem still holds.

**Acknowledgments** This research has been partially supported by the Spanish Ministry of Economy and Competitiveness through the Grant MTM2013-44212-P, the Principality of Asturias/FEDER Grant GRUPIN14-101, Grant C16/15/068 of International Funds KU Leuven and IAP research network grant nr. P7/06 of the Belgian government. Their support is gratefully acknowledged.

# References

1. Beaton AE, Tukey JW (1974) The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 16:147–185
2. Colubi A, González-Rodríguez G (2015) Fuzziness in data analysis: towards accuracy and robustness. Fuzzy Sets Syst 281:260–271
3. Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum K, Hodges JL Jr (eds) A Festschrift for Erich L. Lehmann, Wadsworth
4. Hampel FR (1971) A general qualitative definition of robustness. Ann Math Stat 42(6):1887–1896
5. Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:383–393
6. Hesketh T, Pryor R, Hesketh B (1988) An application of a computerized fuzzy graphic rating scale to the psychological measurement of individual differences. Int J Man-Mach Stud 29:21–35
7. Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101
8. Huber PJ (1981) Robust statistics. Wiley, New York
9. Kim JS, Scott CD (2012) Robust kernel density estimation. J Mach Learn Res 13:2529–2565
10. Puri ML, Ralescu DA (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
11. Sinova B, Gil MA, Colubi A, Van Aelst S (2012) The median of a random fuzzy number. The 1-norm distance approach. Fuzzy Sets Syst 200:99–115
12. Sinova B, Gil MA, Van Aelst S. M-estimates of location for the robust central tendency of fuzzy data. IEEE Trans Fuzzy Syst. Accepted
13. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inf Sci 179:3964–3972

# Technical Gestures Recognition by Set-Valued Hidden Markov Models with Prior Knowledge

Yann Soullard, Alessandro Antonucci and Sébastien Destercke

**Abstract** Hidden Markov models are popular tools for gesture recognition. Once the generative processes of gestures have been identified, an observation sequence is usually classified as the gesture having the highest likelihood, thus ignoring possible prior information. In this paper, we consider two potential improvements of such methods: the inclusion of prior information, and the possibility of considering convex sets of probabilities (in the likelihoods and the prior) to infer imprecise, but more reliable, predictions when information is insufficient. We apply the proposed approach to technical gestures, typically characterized by severe class imbalance. By modelling such imbalances as a prior information, we achieve more accurate results, while the imprecise quantification is shown to produce more reliable estimates.

## 1 Introduction

In this paper we are concerned with classification tasks where one wants to identify gestures (a popular computer vision task [4]) as well as errors in incorrectly executed gestures. We assume the possible gestures belong to a set $\mathcal{C} := \{c_1, \dots, c_M\}$ and denote as $C$ the variable taking values in $\mathcal{C}$. A gesture recognition algorithm then aims at assigning the correct value $c^* \in \mathcal{C}$ to a given sequence. With few exceptions [5], gestures are regarded as multivariate time series, say $(\boldsymbol{o}_1, \dots, \boldsymbol{o}_T)$, with $\boldsymbol{o}_t \in \mathbb{R}^F$ the joint observation of the $F$ features extracted from the $t$-th frame, for each

Y. Soullard (✉) · S. Destercke
Heudiasyc lab, CNRS UMR 7253 Heudiasyc, CS 60 319, 60 203, Université de technologie de, Compiègne cedex, France
e-mail: yann.soullard@hds.utc.fr

S. Destercke
e-mail: sebastien.destercke@hds.utc.fr

A. Antonucci
Istituto Dalle Molle di Studi Sull'Intelligenza Artificiale, Manno-Lugano, Switzerland
e-mail: alessandro@idsia.ch

Y. Soullard · S. Destercke
Sorbonne University, Paris, France

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_56

$t = 1, \ldots, T$. Technical gestures are quite specific, as they are based on particular movements, they require specific skills and they should be executed with a high level of precision. Examples of technical gestures can be found in many domains such as sport (e.g., the forehand of a tennis player), manufacturing (e.g., doing a welding), or handicraft (e.g., the movements of a potter), just to cite a few.

Technical gestures are confronted with specific problems. First, due to the fact that most learning data have to be collected from experts (e.g., if in a later employee training stage, we want to recognize well and badly performed gestures), the obtained data sets are typically small and imbalanced. Those data can also be quite noisy, as measurements are often performed in working environments. Also, when the recognition model is used to decide if a task or a gesture has been performed correctly, a recognition error might have a significant economic impact (e.g. the manufacturing of a defective part or an interruption in the production line). This is why considering tools able to account for this imbalance or this lack of data is important.

Hidden Markov Models (HMMs, [9]) are probabilistic graphical models that can easily cope with multivariate time series, and are therefore often used for gesture recognition [2, 6]. As they are generative models usually trained with maximum-likelihood estimates, HMMs are less prone to over-fitting than their discriminative counterparts [10]. However, they can still suffer from bad parameters estimation when the training examples do not fit well the true data distribution [3]. To gain reliability in the learning, a recent paper [1] proposed a set-valued quantification of the HMM parameters inspired by the theory of imprecise probabilities, for which polynomial-time inference algorithms have been also developed [7]. With those imprecise HMMs, evidential information might not be sufficient to unequivocally recognize the performed gesture, and sets of candidate gestures might be obtained instead. Section 2 contains background information about imprecise methods and HMMs.

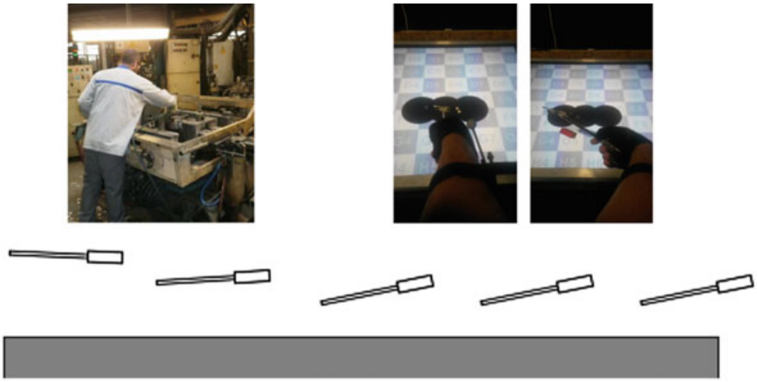


**Fig. 1** Pictures of mold cleaning in a work environment (*top left*) and in the experimental station of a virtual environment (*top right*). Expected positions and inclinations of a blower during a technical gesture with a movement from the *right* to the *left* (*bottom*)

Such approaches take care of the limited amount of available data, while the imbalances over the classes (a typical issue for data of this kind) are neglected by implicitly assuming a uniform marginal distribution over the gestures. The main methodological contribution of this paper, explained in Sect. 3, is a procedure to add prior information about the classes, that can itself be imprecise and represented as a convex set of probability mass functions. The methodology is validated in Sect. 4 on technical gestures performed in an aluminum foundry. This real-world application is part of a training system in a virtual environment for tasks related to mold cleaning (Fig. 1).

## 2 Background

**Imprecise Probability**. Let $C$ denote the class variable associated to the gesture and $\mathcal{C}$ the $M$ possible values. If the uncertainty about $C$ is described by a probability mass function $P$, the task of deciding the actual value of $C$, assuming zero/one losses, returns:

$$c_P^* := \arg\max_{c \in \mathcal{C}} P(c)\,. \tag{1}$$

In many cases single probability mass functions might be unable to provide a reliable uncertainty model. Assume for instance that, among three possible gestures, an expert is telling us that $c_1$ is at least as probable as $c_2$, which is in turn at least as probable as $c_3$. Deciding that $P(C) = [0.7, 0.2, 0.1]$ is a better model than $P'(C) = [0.6, 0.3, 0.1]$ from this information alone is questionable. In such situations, *credal sets*, i.e., closed convex sets of probability mass functions, can offer a more cautious, hence reliable, uncertainty model. In our case, a credal set over $\mathcal{C}$, denoted $K(C)$, will be specified by a finite number of linear constraints, or equivalently by its (finite) set of extreme points. In the expert example with three gestures, we can consider the credal set $K(C)$ defined by the constraints $P(c_1) \geq P(c_2) \geq P(c_3)$, together with non-negativity and normalization, or equivalently, by listing the extreme points $P_1(C) = [1, 0, 0]$, $P_2(C) = [\frac{1}{2}, \frac{1}{2}, 0]$, and $P_3(C) = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ (Fig. 2). The general-

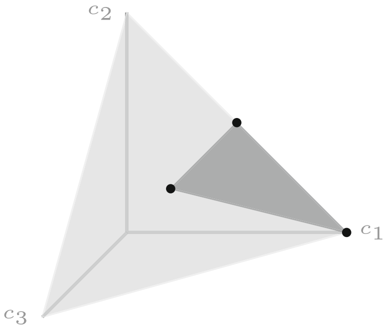


**Fig. 2** A credal set modeling uncertainty about a gesture with three options

ization of Eq. (1) to credal sets can be achieved in many ways. Here we consider the *maximality* criterion, which returns the following sets of optimal classes:

$$\mathcal{C}_K^* := \{c' \in \mathcal{C} : \nexists\, c'' \in \mathcal{C}\,\text{s.t.}\ P(c'') > P(c') \forall\, P(C) \in K(C)\}\,. \tag{2}$$

Non-optimal classes are therefore those such that, for each element of the credal set, there is another class with strictly higher probability.

**Hidden Markov Models (HMMs)**. HMMs [9] are popular probabilistic descriptions of time series with many applications in speech recognition and computer vision, to name but a few. HMMs assume the observation $\boldsymbol{O}_t$ is generated by a paired *state* variable $X_t$, for each $t = 1, \ldots, T$, with $T$ the length of the sequence. State variables are in turn assumed to be generated by a Markov chain process. All state variables take their values from a space $\mathcal{X}$ of cardinality $N$. An HMM specification comprises an initial state probability mass function $P(X_1)$, a $N \times N$ state transition probability matrix $P(X_{t+1}|X_t)$, and a (usually normal) distribution for each observation with mean and covariance indexed by the corresponding state, say $\boldsymbol{\mu}(X_t)$ and $\boldsymbol{\sigma}(X_t)$. We consider *stationary* models with the values of the parameters independent of $t$. HMMs give a compact specification of the joint density:

$$P(x_1, \ldots, x_T, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_T) := P(x_1) \prod_{t=1}^{T-1} P(x_{t+1}|x_t) \prod_{t=1}^{T} \mathcal{N}_{\boldsymbol{\sigma}(x_t)}^{\boldsymbol{\mu}(x_t)}(\boldsymbol{o}_t)\,. \tag{3}$$

By marginalizing the states in Eq. (3) we obtain the *likelihood* of a sequence $P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T)$. This can be achieved in $O(TN^2)$ time by a message propagation algorithm [9]. HMMs are trained using an Expectation-Maximization approach, the Baum-Welch algorithm, detecting a local maximum of the likelihood defined by the joint probabilities of the training sequences and of their classes. Classification can then be achieved by: (i) training a HMM per class; and then (ii) assigning to a test sequence $(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T)$ the class associated to the HMM giving the highest likelihood to the sequence, i.e.,

$$c^* := \arg\max_{c \in \mathcal{C}} P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c)\,, \tag{4}$$

where notation $P(\ldots|c)$ is used for the density corresponding to the HMM associated to class $c$. Here no prior probabilities over the classes are supposed to be available, i.e., a uniform distribution over them is implicitly assumed.

As Baum-Welch estimates might be unreliable, for instance when using few data or short sequences, imprecise probabilities have been proposed to mitigate this unreliability in the HMM quantification [1]. An HMM with imprecise parameters can be learned from a sequence by combining the Baum-Welch algorithm with the *imprecise Dirichlet model* (IDM, [11]). In this model, $P(X_1)$ is replaced by a credal set $K(X_1)$ and $P(X_{t+1}|x_t)$ with $K(X_{t+1}|x_t)$ for each $x_t$. As shown in [7], the bounds $[\underline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c), \overline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c)]$ of the likelihood with respect to those credal sets can be computed with the same time complexity as the precise computation. The classification scheme in Eq. (4) can then be extended to set-valued HMMs by comparing the likelihood intervals and then deciding the optimal ones as in Eq. (2).

## 3 HMM-Based Classification with Prior Knowledge

If prior knowledge about the classes is available in the form of a mass function $P(C)$, the likelihood-based classification scheme in Eq. (4) becomes:

$$c^* = \arg\max_{c\in\mathcal{C}} P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c) \cdot P(c)\,, \tag{5}$$

which corresponds to a comparison of the posterior probabilities

$$P(c|\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T) \propto P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c) \cdot P(c)\,. \tag{6}$$

A proper assessment of the prior mass function is clearly crucial in this Bayesian framework. Yet, the elicitation of qualitative or quantitative expert prior knowledge suffers from the same issues discussed in Sect. 2, and a credal set $K(C)$ might offer a more reliable model of the prior knowledge about $C$. We therefore consider a twofold generalization of Eq. (5) to imprecise probabilities in which $P(C)$ is replaced by a credal set $K(C)$, and the sequence likelihoods $P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c)$ are replaced by their lower/upper bounds learned from the training data. The optimal classes can be therefore obtained by applying the criterion in Eq. (2) to the, imprecisely specified, posterior probabilities in Eq. (6). To achieve that in practice, given two classes $c', c'' \in \mathcal{C}$, we evaluate whether the posterior probability for $c''$ is always greater than that of $c'$, i.e.,

$$\min_{\substack{P(C)\in K(C)\\ P(\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T|C)\in[\underline{P}(\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T|C),\overline{P}(\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T|C)]}} \frac{P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c'') \cdot P(c'')}{P(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c') \cdot P(c')} > 1\,. \tag{7}$$

where we assume the denominator strictly positive. If the above inequality is satisfied, class $c'$ is removed from the set of optimal labels. The set of optimal options $\mathcal{C}_K^*$ is obtained by iterating the test in Eq. (7) for any pair of classes, and removing from $\mathcal{C}$ the dominated options. The optimization with respect to the imprecisely specified likelihoods is trivial and allows to rewrite Eq. (7) as follows:

$$\min_{P(C)\in K(C)} \frac{\underline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c'') \cdot P(c'')}{\overline{P}(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T|c') \cdot P(c')} > 1\,. \tag{8}$$

As $K(C)$ can be expressed by linear constraints, the task in Eq. (8) is a linear-fractional task, which can be reduced to a linear program and solved in polynomial time w.r.t. the number of classes $M$ by a linear solver.

## 4 Empirical Validation

We test the proposed approach on six technical gesture data sets (Table 1). The TG and TGE datasets refer respectively to classification of types of gestures and types of errors (for specific gestures). The gestures are performed in an aluminium foundry and refer to a workstation where a technician cleans a mold (Fig. 1). The technician performs several tasks with different tools such as a compressed-air blower, a scraper and a pistol. Motion capture is performed by markers attached to the tools and the user's body. Markers are tracked by infrared cameras and, at each time frame, 3D positions and orientations are extracted. Such raw features may not directly provide a good modelling of the gesture. Following [8], we compute high-level features such as velocities, pairwise distances and angles to enrich the description.

To train HMMs as in Eq. (3), we run the Baum-Welch algorithm with a maximum of 25 iterations before convergence and three states for the hidden variables (i.e., $N = 3$). For the imprecise quantification we set $s = 4$ for the parameter determining the imprecision level (in term of missing observations) in the IDM. The *accuracy* (i.e., the percentage of properly classified gestures) describes the performance of the precise classifiers. We say that an imprecise classifier is *indeterminate* when more than one class is returned as output. To characterize the output of an imprecise classifier we use its *determinacy* (i.e., percentage of determinate outputs) and *output size* (i.e., average number of classes in output when indeterminate). The performance is described in terms of *single accuracy* (i.e., accuracy when the output is determinate) and *set accuracy* (i.e., percentage of indeterminate outputs including the true class). For a direct comparison with precise classifiers we compare the accuracy with the $u_{80}$ utility-based measure. This is basically a positive correction (namely $1.2(q-1)/q^2$), advocated in [12], of a discounted accuracy giving $1/q$ to a classifier returning $q$ options if one of them is correct, and zero otherwise.

The proposed method is intended to achieve robustness when coping with small datasets. Accordingly, we adopt a (fivefold) cross validation scheme with one fold for training, and the rest for testing. In Fig. 3, we compare the accuracies of the approaches based on the likelihood (Eq. (4)) and the posterior (Eq. (6)) with the $u_{80}$ for the imprecise posterior. The precise prior is obtained from the distribution

**Table 1** Number of features, classes, and samples per class in the benchmark

| Dataset | F | M | Samples for $c_1/\dots/c_m$ |
|---|---|---|---|
| $TG_1$ | 15 | 4 | 320/160/224/287 |
| $TG_2$ | 15 | 4 | 192/320/256/287 |
| $TG_3$ | 18 | 4 | 100/100/40/20 |
| $TGE_1$ | 19 | 4 | 57/36/45/33 |
| $TGE_2$ | 4 | 3 | 15/30/20 |
| $TGE_3$ | 4 | 3 | 20/10/15 |

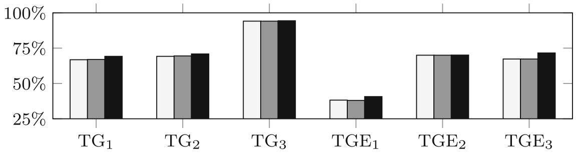


**Fig. 3** Accuracies of the likelihood (*white*) and posterior (*gray*) comparison against the $u_{80}$ of the imprecise posterior (*black*)

**Table 2** Performance of the classifier in the precise and imprecise posterior case

| Dataset | Precise accuracy (%) | Single accuracy (%) | Set accuracy (%) | Determinacy (%) | Output size |
|---|---|---|---|---|---|
| TG$_1$ | 67.3 | 70.3 | 80.8 | 93.0 | 2.1 |
| TG$_2$ | 69.7 | 71.5 | 78.8 | 93.7 | 2.1 |
| TG$_3$ | 94.7 | 95.0 | 100.0 | 97.9 | 2.0 |
| TGE$_1$ | 38.0 | 40.7 | 58.7 | 96.2 | 2.0 |
| TGE$_2$ | 70.0 | 71.1 | 76.7 | 94.6 | 2.0 |
| TGE$_3$ | 67.3 | 71.1 | 100.0 | 93.6 | 2.0 |

over the classes of the training data. The prior credal set is similarly obtained by the IDM ($s = 4$). Introducing the prior has a positive effect which is only modest in the precise case and more notable in the imprecise case. A deeper analysis of the imprecise model based on the posterior is in Table 2. Remarkably, the classifier achieves high determinacies and, when indeterminate, only two classes are typically returned. The single accuracies are higher than the accuracies of the precise models (i.e., when determinate the imprecise classifier outperforms the precise methods). Finally, on two datasets, when indeterminate the imprecise classifier returns always two classes and one of them is always the correct one.

## 5 Conclusions and Outlooks

A new classification algorithm for multivariate time series is proposed. The sequences are described by HMMs, and the likelihoods returned by these models are combined with a prior distribution over the classes. A robust modeling based on an imprecise-probabilistic quantification of the HMM parameters and the prior is shown to produce more reliable classification performance, without compromising the computational efficiency. Such an approach allows to deal with small and imbalanced datasets. We obtain a set of predicted labels when the information is not sufficient to recognize the performed gesture. An application to technical gesture recognition in an industrial context is reported. As future work, we want to apply our approach to sequences of gestures, by also achieving a segmentation of the various gestures.

**Acknowledgments** This work is founded by the European Union and the French region Picardie. Europe acts in Picardie with the *European Regional Development Fund (ERDF)*. This work is supported by the ANR UML-net project, grant ANR-14-CE24-0026 of the French *Agence Nationale de la Recherche (ANR)*.

## References

1. Antonucci A, de Rosa R, Giusti A, Cuzzolin F (2015) Robust classification of multivariate time series by imprecise hidden Markov models. Int J Approx Reason 56(B):249–263
2. Bevilacqua F, Zamborlin B, Sypniewski A, Schnell N, Guédy F, Rasamimanana N (2010) Continuous Realtime Gesture Following and Recognition. In: Gesture in embodied communication and human-computer interaction: 8th international gesture workshop, GW 2009, Revised Selected Papers. Springer, pp 73–84
3. Bouchard G, Triggs B (2004) The tradeoff between generative and discriminative classifiers. In: International symposium on computational statistics, pp 721–728
4. Chaudhary A, Raheja JL, Das K, Raheja S (2011) Intelligent approaches to interact with machines using hand gesture recognition in natural way: a survey. Int J Comput Sci Eng Surv 2(1):122–133
5. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247–2253
6. Liu K, Chen C, Jafari R, Kehtarnavaz N (2014) Multi-HMM classification for hand gesture recognition using two differing modality sensors. In: Circuits and systems conference (DCAS). IEEE, pp 1–4
7. Mauá DD, Antonucci A, de Campos CP (2015) Hidden Markov models with set-valued parameters. Neurocomputing 180:94–107
8. Neverova N, Wolf C, Taylor GW, Nebout F (2014) Multi-scale deep learning for gesture detection and localization. In: ECCV workshop on looking at people
9. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
10. Soullard Y, Saveski M, Artières T (2014) Joint semi-supervised learning of hidden conditional random fields and hidden Markov models. Pattern Recogn Lett 37:161–171
11. Walley P (1996) Inferences from multinomial data: learning about a bag of marbles. J R Stat Soc B 58(1):3–57
12. Zaffalon M, Corani G, Mauá DD (2012) Evaluating credal classifiers by utility-discounted predictive accuracy. Int J Approx Reason 53(8):1282–1301

# Time Series Modeling Based on Fuzzy Transform

**Luciano Stefanini, Laerte Sorini and Maria Letizia Guerra**

**Abstract** It is well known that smoothing is applied to better see patterns and underlying trends in time series. In fact, to smooth a data set means to create an approximating function that attempts to capture important features in the data, while leaving out noises. In this paper we choose, as an approximation function, the inverse fuzzy transform (introduced by Perfilieva in Fuzzy Sets Syst 157:993–1023, 2006 [3]) that is based on fuzzy partitioning of a closed real interval into fuzzy subsets. The empirical distribution we introduce can be characterized by its expectiles in a similar way as it is characterized by quantiles.

## 1 Basic Mathematical Tools

All the main following results come from the seminal paper [3] and from the papers [7, 8] and [11]. A fuzzy partition $(P, A)$ for a real compact interval $[a, b]$ is build by a decomposition $P = \{a = x_1 < x_2 < \cdots < x_n = b\}$ of $[a, b]$ into $n - 1$ subintervals $[x_{k-1}, x_k]$, $k = 2, \ldots, n$ and by a family $A = \{A_1, A_2, \ldots, A_n\}$ of $n$ fuzzy numbers identified by the membership functions $A_1(x), A_2(x), \ldots, A_n(x)$ for $x \in [a, b]$ satisfying some properties:

1. *each* $A_k : [a, b] \longrightarrow [0, 1]$ *is continuous with* $A_k(x_k) = 1$, $A_k(x) = 0$ *for* $x \notin [x_{k-1}, x_{k+1}]$*;*
2. *for* $k = 2, 3, \ldots, n - 1$, $A_k$ *is increasing on* $[x_{k-1}, x_k]$ *and decreasing on* $[x_k, x_{k+1}]$*;* $A_1$ *is decreasing on* $[a, x_2]$*;* $A_n$ *is increasing on* $[x_{n-1}, b]$*;*
3. *for all* $x \in [a, b]$ *the following partition-of-unity condition holds*

L. Stefanini · L. Sorini
Department of Economics, Society and Politics, University of Urbino, Urbino, Italy
e-mail: luciano.stefanini@uniurb.it

L. Sorini
e-mail: laerte.sorini@uniurb.it

M.L. Guerra (✉)
Department of Mathematics, University of Bologna, Bologna, Italy
e-mail: mletizia.guerra@unibo.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_57

$$\sum_{k=1}^{n} A_k(x) = 1.$$

Given a continuous function $f : [a, b] \longrightarrow R$ and a fuzzy partition $(P, A)$ of $[a, b]$, the direct Fuzzy transform (F-transform) of $f$ with respect to $(P, A)$ is the following $n$-tuple of real numbers $F = (F_1, F_2, \ldots, F_n)^T$ where

$$F_k = \frac{\int_a^b f(x)A_k(x)dx}{\int_a^b A_k(x)dx} = \frac{\int_{x_{k-1}}^{x_{k+1}} f(x)A_k(x)dx}{\int_{x_{k-1}}^{x_{k+1}} A_k(x)dx}, \quad k = 1, 2, \ldots, n \tag{1}$$

Given the direct fuzzy transform $(F_1, F_2, \ldots, F_n)^T$ of a continuous function $f : [a, b] \longrightarrow R$ on a fuzzy partition $(P, A)$, the inverse F-transform (iF-transform) is the continuous function $\widehat{f_{\mathbf{F}}} : [a, b] \longrightarrow R$ given by

$$\widehat{f_{\mathbf{F}}}(x) = \sum_{k=1}^{n} F_k A_k(x)\, for\, x \in [a, b]. \tag{2}$$

The inverse F-transform function $\widehat{f_{\mathbf{F}}} : [a, b] \longrightarrow R$ is an approximating function of $f$ on $[a, b]$.

If $f : [a, b] \longrightarrow R$ is a continuous function then, for any positive real $\varepsilon$, there exists a fuzzy partition $(P_\varepsilon, A_\varepsilon)$ such that the associated F-transform $F_\varepsilon = (F_{1,\varepsilon}, F_{2,\varepsilon}, \ldots, F_{n_\varepsilon,\varepsilon})^T$ and the corresponding iF-transform $\widehat{f_{\mathbf{F}_\varepsilon}} : [a, b] \longrightarrow R$ satisfies

$$\left| f(x) - \widehat{f_{\mathbf{F}_\varepsilon}}(x) \right| < \varepsilon\, for\, all\, x \in [a, b].$$

The most important property is that:

$$\int_a^b f(x)dx = \int_a^b \widehat{f_{\mathbf{F}}}(x)dx$$

implying the existence of an accurate smoothing technique that preserves the areas.

We can then define an $r$-partition in the following way.

Let $r \geq 1$ be a fixed integer number; a fuzzy $r$-partition of $[a, b]$ is given by a pair $(P, A^{(r)})$ where $P = \{a = x_1 < \cdots < x_n = b\}$ is a decomposition of $[a, b]$, and $A^{(r)}$ is a family of $n + 2r - 2$ continuous, normal, convex fuzzy numbers

$$\mathbb{A}^{(r)} = \{A_k^{(r)} : [a, b] \longrightarrow [0, 1] | k = -r + 2, \ldots, n + r - 1\}$$

such that

a. for $k = 1, 2, \ldots, n$, $A_k^{(r)}$ is a continuous fuzzy number with $A_k^{(r)}(x_k) = 1$ and $A_k^{(r)}(x) = 0$ for $x \notin [x_{k-r}, x_{k+r}]$;
b. for $k = 1, 2, \ldots, n$, $A_k^{(r)}$ is increasing on $[x_{k-r}, x_k]$ and decreasing on $[x_k, x_{k+r}]$;

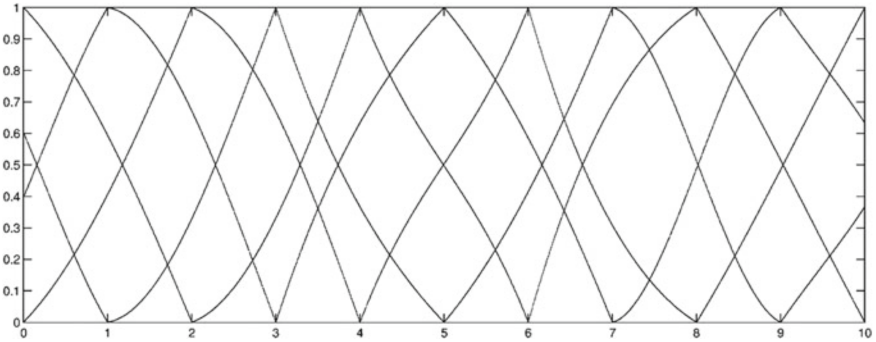


**Fig. 1** Generalized parametric fuzzy partition $r = 2$

c. for $k = -r + 2, \ldots, 0$, $A_k^{(r)}$ is decreasing on $[x_k, x_{k+r}]$;
d. for $k = n + 1, \ldots, n + r - 1$, $A_k^{(r)}$ is increasing on $[x_{k-r}, x_k]$;
e. for all $x \in [a, b]$, the partition-of-$r$ condition holds $\sum_{k=-r+2}^{n+r-1} A_k^{(r)}(x) = r$.

The integer $r \geq 1$ will be called the bandwidth of the partition $(P, A^{(r)})$ and the effect of the smoothing is higher when the bandwidth is greater than 1, in fact for $r = 1$ the smoothing has no effect.

In the same way, when $r = 2$, we obtain the fuzzy 2-partition $(P, A^{(2)})$ that is shown in the following figure when we take under consideration 2 intervals before and 2 intervals after (Fig. 1):

At this level, the direct F$^{(r)}$-transform based on the generalized fuzzy r-partition $(P, A^{(r)})$ can be introduced and it is defined by the vector $F^{(r)} = (F_1^{(r)}, F_2^{(r)}, \ldots, F_n^{(r)})^T$, where

$$F_k^{(r)} = \frac{1}{I_k^{(r)}} \int_a^b f(x) A_k^{(r)}(x) dx \text{ for } k = 1, 2, \ldots, n \tag{3}$$

$$I_k^{(r)} = \int_a^b A_k^{(r)}(x) dx. \tag{4}$$

The $iF^{(r)}$-transform function (of bandwidth $r$) is

$$\widehat{f}^{(r)}(x) = \frac{1}{r} \sum_{k=1}^{n} F_k^{(r)} A_k^{(r)}(x). \tag{5}$$

On the other hand, the $iF^{(r)}$-transform function $\widehat{f}^{(r)}(x)$ has the structure of a moving average of the values $\{F_j^{(r)}, j = 1, \ldots, n\}$; when $F_k^{(r)} = 0$ if $k < 1$ or $k > n$, we have

$$\widehat{f}^{(r)}(x) = \frac{1}{r} \sum_{j=k-r}^{k+r} F_j^{(r)} A_j^{(r)}(x), \tag{6}$$

i.e., a weighted average of $F_{k-r}^{(r)}, \ldots, F_k^{(r)}, \ldots, F_{k+r}^{(r)}$ with weights $\frac{A_{k-r}^{(r)}(x)}{r}, \ldots, \frac{A_k^{(r)}(x)}{r}, \ldots, \frac{A_{k+r}^{(r)}(x)}{r}$.

The main properties of the $F^{(r)}$-transform are analogues to the properties of the standard $F$-transform.

## 2 Quantile and Expectile Smoothing

Consider a real-valued random variable $\xi$; a given $r$-quantile $\xi(r)$ (where $r$ plays now a different role) means that the probability that an observation is less than $\xi(r)$ is $r$, with $r \in ]0, 1[$. Given a set of $T$ observations $y_t, t = 1, \ldots, T$, the sample quantile $\overline{\xi}(r)$ can be obtained as the solution to minimize the function

$$S_r(\xi) = \sum_{y_t < \xi} (1 - r)(\xi - y_t) + \sum_{y_t \geq \xi} r(y_t - \xi) \geq 0.$$

If $r = \frac{1}{2}$, then the $r$-quantile gives the median of the (empirical) distribution, i.e. the minimizer of the functional

$$S_{1/2}(\xi) = \sum_{t=1}^{T} |y_t - \xi| \geq 0.$$

However,.a distribution can be also characterized by its expectiles that minimize a quadratic functional and so they work like mean values or by its quantiles that minimize the absolute value of the difference.

The expectiles are defined in a similar way as for quantiles except that they are defined by tail expectations and by using the mean instead of the median. Quantiles have a strong intuitive appeal, but expectiles are easier to compute and expectile approach is probably more interesting because the related operator is differentiable while for the quantile this is not true.

The efficiency of expectiles is clear when smoothing small data sets; in fact, least asymmetrically weighted squares make use of the distance to data points in estimating a curve. Quantile smoothing only knows whether an observation is below or above the curve while expectiles are much more sensitive to outliers than quantiles.

The sample expectile $\mu(r)$ can be obtained as the solution to minimize the following function

$$S_r(\mu) = \sum_{\substack{t=1 \\ x_t < \mu}}^{T} (1 - r)(x_t - \mu)^2 + \sum_{\substack{t=1 \\ x_t > \mu}}^{T} r(x_t - \mu)^2.$$

If $r = \frac{1}{2}$ we obtain the mean value $\mu_e$ of the observations

$$\mu_e = \arg\min_{\mu} S_{\frac{1}{2}}(\mu) = \frac{1}{2}\sum_{t=1}^{T}(x_t - \mu)^2$$
$$\mu_e = \frac{1}{T}\sum_{t=1}^{T} x_t$$

The expectile F-transform, for a fixed generalized fuzzy r-partition $(P, A^{(r)})$ and for a given value of $r \in ]0, 1]$, can be defined as the minimizer of the following operators, for $k = 1, \ldots, n$,

$$\Phi_{k,r}(F) = \int_a^b w_r(x)\,(f(x) - F)^2 A_k^{(r)}(x)\,dx$$
$$\text{where}$$
$$w_r(x) = \begin{cases} r & \text{if } f(x) \le F \\ 1 - r & \text{if } f(x) > F \end{cases};$$

The quantile F-transform, for a fixed generalized fuzzy r-partition $(P, A^{(r)})$ and for a given value of $r \in ]0, 1]$, can be defined as the minimizer of the following operators, for $k = 1, \ldots, n$,

$$\Psi_{k,r}(F) = \int_a^b w_r(x)\,|f(x) - F|\,A_k^{(r)}(x)\,dx$$

If $\alpha = 1$ we obtain $\Phi_{k,0.5}(F)$.

The minimization of $\Phi_{k,\alpha}^{-}(F)$ and $\Phi_{k,\alpha}^{+}(F)$ produces, respectively $F_{k,\alpha}^{-}$ and $F_{k,\alpha}^{+}$ so that $\left[F_{k,\alpha}^{-}, F_{k,\alpha}^{+}\right]$ is the $\alpha$-cut of $F_k$.

As a consequence, the *iF*-transform of $f$ is fuzzified by:

$$\widehat{f}(x) = \frac{1}{r}\sum_{k=1}^{n} F_k A_k^{(r)}(x)$$

with the corresponding *α-cuts* expressed as:

$$\left[\widehat{f}(x)\right]_\alpha = \left[\widehat{f}_\alpha^{-}(x), \widehat{f}_\alpha^{+}(x)\right]_\alpha \tag{7}$$
$$= \left[\frac{1}{r}\sum_{k=1}^{n} F_{k,\alpha}^{-} A_k^{(r)}(x), \frac{1}{r}\sum_{k=1}^{n} F_{k,\alpha}^{+} A_k^{(r)}(x)\right]$$

When $\alpha = 1$ we obtain the standard $F$-transform and the corresponding *iF* transform.

The discrete case can be handled in a similar way as for the standard discrete F-transform. The expectiles, in the discrete case, are obtained by minimizing the following functions:

$$\Phi^{-}_{k,\alpha}(F) = \sum_{k=1}^{m} w^{-}_{\alpha}(t_i)\,(f(t_i) - F)^2 A^{(r)}_{k}(t_i)$$

where

$$w^{-}_{\alpha}(t_i) = \begin{cases} \frac{\alpha}{2} & \text{if } f(t_i) \le F \\ 1 - \frac{\alpha}{2} & \text{if } f(t_i) > F \end{cases}$$

$$\Phi^{+}_{k,\alpha}(F) = \sum_{k=1}^{m} w^{+}_{\alpha}(t_i)\,(f(t_i) - F)^2 A^{(r)}_{k}(t_i)$$

where

$$w^{+}_{\alpha}(t_i) = \begin{cases} 1 - \frac{\alpha}{2} & \text{if } f(t_i) \le F \\ \frac{\alpha}{2} & \text{if } f(t_i) > F \end{cases}$$

Consider that, for fixed values $w^{\pm}_{\alpha}(t_i) = w_i$, the minimizer $F_{k,\alpha}$ is obtained by

$$F_{k,\alpha} = \frac{\sum_{k=1}^{m} w_i f(t_i)\,A^{(r)}_{k}(t_i)}{\sum_{k=1}^{m} w_i A^{(r)}_{k}(t_i)}, k = 1, \ldots, n$$

and an iterative procedure can be adopted, similar to the one described above.

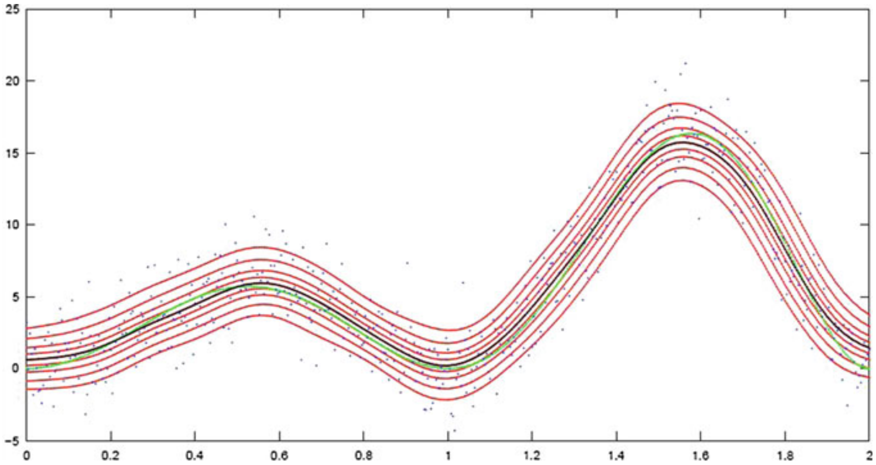


**Fig. 2** $\alpha$-cuts of a fuzzy-valued function by F-transform $(m = 501, n = 101, r = 6)$ $\alpha = 0.01, 0.25, 0.5, 0.75, 1.0$

In Fig. 2 the performance of the expectile smoothing approximation with F-transform is illustrated when $f(t_i) = 5e^{-0.5t_i^2}\sin^2(\pi t_i) + 2z_i,\ \ t_i \in [0, 2],\ \ i = 1, \ldots, m$, where $z_i \in N(0, 1)$.

The data are represented by points and 9 curves are generated, corresponding to the values of $\alpha = 0.01, 0.25, 0.5, 0.75, 1.0$; it is to be remarked that for any value of $\alpha \in ]0, 1]$ we can obtain the $\alpha-$cut $\left[F^{-}_{k,\alpha}, F^{+}_{k,\alpha}\right]$ of $F_k$, $k = 1, 2, \ldots, n$. The curves are then constructed by inverse F-transform.

## 3 Examples

In order to show how the F-transform can be used for expectile smoothing, we apply the proposed estimation on one financial time series. The number $n$ of subintervals in the fuzzy partition $(P, A^{(r)})$ are approximately $\frac{m}{5}$ and the bandwidth $r$ is estimated by generalized cross validation. In all cases, for simplicity, the basic functions $A_k(x)$, defined on the intervals $[x_{k-r}, x_{k+r}]$, are obtained by translating and rescaling the same symmetric triangular fuzzy number $T_0$, defined on $[-1, 1]$ and centered at the origin, with membership

$$T_0(t) = \begin{cases} 1 + t \text{ if } t \in [-1, 0] \\ 1 - t \text{ if } t \in [0, 1] \\ 0 \qquad \text{otherwise} \end{cases}.$$

The time series in Fig. 3 is the daily London Gold Fixing, the usual benchmark for the gold price; it also provides a published benchmark price that is widely used

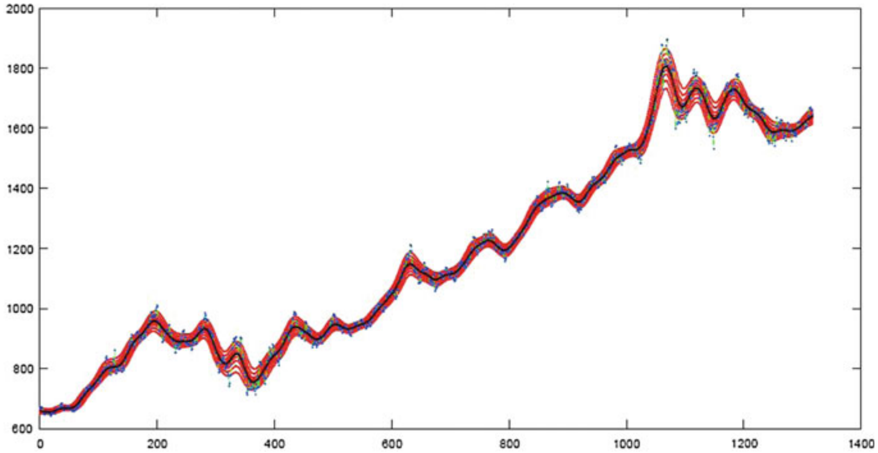


**Fig. 3** $\alpha$-cuts of a fuzzy-valued function by F-transform $(m = 1317, n = 250, r = 3)$ $\alpha = 0.01, 0.25, 0.5, 0.75, 1.0$

as a pricing medium by producers, consumers, investors and central banks. The $m = 1317$ observations cover the period from June 2007 to August 2012.

The introduced smoothing technique may represent a good alternative to the most popular ones, for example LOWESS (Locally Weighted Scatterplot Smoothing), because it always produces monotonic behaviors and the algorithmic implementation is simple while the integral is preserved.

In addition, using an appropriate (generalized) fuzzy partition, the $\alpha$-cuts $\left[F^-_{k,\alpha}, F^+_{k,\alpha}\right]$ of $F_k$ have the same smoothing property inherited from F-transform, with a "degree of smoothness" depending on the bandwidth of the partition.

The preliminary results encourage to further work in the study and applications of F-transform as a tool to obtain a fuzzy-valued interpretation of a time series.

## References

1. Di Martino F, Loia V, Sessa S (2010) Fuzzy transforms method and attribute dependency in data analysis. Inf Sci 180:493–505
2. Guerra ML, Stefanini L (2013) Expectile smoothing of time series using F-transform. In: Conference of the European society for fuzzy logic and technology, advances in intelligent systems research, vol 32, pp 559–564. ISBN: 978-162993219-4
3. Perfilieva I (2006) Fuzzy transforms: theory and applications. Fuzzy Sets Syst 157:993–1023
4. Perfilieva I (2006) Fuzzy transforms and their applications to data compression. In: Bloch I et al (eds) Fuzzy logic and applications, LNAI 3849. Springer, pp 19–31
5. Perfilieva I, Novak V, Dvorak A (2008) Fuzzy transform in the analysis of data. Int J Approx Reason 48:36–46
6. Schnabel SK, Eilers PHC (2009) Optimal expectile smoothing. Comput Stat Data Anal 53:4168–4177
7. Stefanini L (2008) Fuzzy transform with parametric LU-fuzzy partitions. In: Ruan D et al (eds) Computational intelligence in decision and control. World Scientific, pp 399–404
8. Stefanini L (2009) Fuzzy transform and smooth functions. In: Carvalho JP, Dubois D, Kaymak U, Sousa JMC (eds) Proceeding of the IFSA-EUSFLAT 2009 conference, Lisbon, July 2009, pp 579–584
9. Stefanini L, Sorini L, Guerra ML (2006) Parametric representations of fuzzy numbers and application to fuzzy calculus. Fuzzy Sets Syst 157:2423–2455
10. Stefanini L, Sorini L, Guerra ML (2009) Fuzzy numbers and fuzzy arithmetic. In: Pedrycz W, Skowron A, Kreynovich V (eds) Handbook of granular computing, Chapter 12. Wiley
11. Stefanini L (2011) F-transform with parametric generalized fuzzy partitions. Fuzzy Sets Syst 180(1):98–120

# Back to "Reasoning"

**Marco Elio Tabacchi and Settimo Termini**

**Abstract** Is rigor always strictly related to precision and accuracy? This is a fundamental question in the realm of Fuzzy Logic; the first instinct would be to answer in the positive, but the question is much more complex than it appears, as true rigor is obtained also by a careful examination of the context, and limiting to a mechanical transfer of techniques, procedures and conceptual attitudes from one domain to another, such as from the pure engineering feats or the ones of mathematical logic to the study of human reasoning, does not guarantee optimal results. Starting from this question, we discuss some implications of going back to the very concept of reasoning as it is used in natural language and in everyday life. Taking into account the presence—from the start—of uncertainty and approximation in one of its possible forms seems to indicate the need of a different approach from the simple extension of tools and concepts from mathematical logic.

## 1 Introduction

Had the format allowed it, a possible, albeit very long, subtitle to this paper could have been: "There are more things in the world of fuzzy logic (with respect to the possibility of picking up relevant aspects of *reasoning*) than in formalized mathematical logics." The previous sentence does not fit well with the role of a subtitle, but can perhaps play a role for clarifying the aims of the present paper. First of all, let us stress that we are using the term *reasoning* to discuss the informal (but rigorous) use of the term. What could be the content and the scope, for instance, of an invitation—when facing a difficult problem—of this kind: *let us discuss about it* (equivalent of *let us reason about it*). We want to stress that the use of *reason* here, as a sort of synonym of *discuss* presents the two following features: (a) The procedure followed along the dialogue

M.E. Tabacchi (✉)
DMI Unipa and INR Demopolis, Palermo, Italy
e-mail: metabacchi@unipa.it; metabacchi@demopolis.it

S. Termini (✉)
DMI UNipa, Palermo, Italy
e-mail: settimo.termini@unipa.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_58

is presumed to be very rigorous without any sloppiness neither in the presentation of the problem nor regarding the argumentation. (b) One does not think that this *piece of reasoning*—i.e., *piece of discussion*—will be formalized, and for the simple fact that no advantage could be envisaged from a possible formalization. We can have rigor without formalization, as is well known: this is what usually happens to human beings in their act of reasoning. Let us outline the plan of the paper. In the following Sect. 2 we discuss the relationship among rigor, precision and accuracy. In Sect. 3, we relate these considerations with the heritage of mathematical logic to any modelling of *reasoning*. In Sect. 4 we ask which notions are really important in a "reasoning context", in which uncertainty, fuzziness, vagueness are important players of the game, looking for meaningful aspects of the real processes of reasoning, embedded by complex constraints which impede too harsh simplifications. We are confident that this *impure* setting is what allows creativity, adaptations and the like, by easily allowing to switch from one context to another one. Conclusions will follow.

## 2 Rigor, Precision and Accuracy

Let us ask a question: "Is rigor always strictly related to precision and accuracy?" One would, perhaps, be induced to immediately answer *Yes*. However, under reflection, it is clear that the situation is, by far, much more complex. True rigor is obtained also by a careful examination of the context in which we move and the mechanical transferring of techniques, procedures and conceptual attitudes from one domain to another one can produce undesirable results. Moreover a local increase of precision and accuracy can imbalance all the system, producing a collapse of the equilibrium among the parts and, as a consequence, worst results. We think that we should take into account a useful lesson which starting, at least, from Aristotle arrives to Karl Popper. A forgotten lesson, we would say. Let us observe that it is surely good to take inspiration from "good practices" and, in particular, to see what is the behaviour of a successful discipline in relation to both precision and rigor, and model our action in another domain accordingly. However we must be careful in not applying the recipe in a disastrous way by a mechanical transferring of the original methodology, guided by just a few rules. The good aspect is to take as a guiding example the fields and disciplines in which a high level of rigor has been obtained. The bad aspect is to force the same methods in an uncritical way to very different domains, something which can produce unpleasant results when there is a very simplified and, in some cases, sloppy use of very beautiful and sophisticated constructions designed for completely different aims. To this aim it would be useful to always recall Aristotle's comment: "It is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits." And, to reinforce what he has in mind, he continues by saying: "it is evidently equally foolish to accept probable reasoning from a mathematician and to demand from a rhetorician scientific proofs." (Nicomachean Ethics, book I Chapter 3, translated by W.D. Ross) The fact is, as Karl Raimund Popper writes, that "both precision and certainty are false ideals. They are impossible to attain, and therefore dangerously misleading if they are uncritically accepted as guides. The quest for precision is analogous to the quest for certainty, and both should be abandoned.” [13, p. 22] Maybe such harsh affirmation of a principle can shock the listener, but what follows underlines the importance of the point: “I do not suggest, of course, that an increase in the precision of, say, a prediction, or even a formulation, may not sometimes be highly desirable. What I do suggest is that it is always undesirable to make an effort to increase *precision for its own sake*—especially linguistic precision—since this usually leads to loss of clarity…” [13, p. 22] (italics ours). His conclusion is that “one should never try to be more precise than the problem situation demands.” [13, p. 22] Somehow this statement is parallel to the previously mentioned Aristotle’s. We shall try with his help to understand better the situation. An increase in the clarity with which we present a problem is always useful and welcome. However this clarity is not always related to an increase in the precision with which we describe parts of our problem. This can be so if this increase in the accuracy of a certain measurements is useful (to be sure that there are no harmful bacteria in a throat, or for choosing between two theories). But in itself increasing accuracy and precision is not a virtue, while the increase in clarity when posing a problem is always desirable. What could be the origin of this desire to look for “precision for its own sake”? What is the origin of this attitude? This is, perhaps, connected to the tendency, denounced above, of transferring in a mechanical way, rigorous approaches which elsewhere have worked well. Maybe this is due to the successes of mathematics in physics and the success of physics when it began to use mathematics to develop the intuitions of the way in which the world works. It seems to us that Popper describes in a clear way a situation very similar to the questions we aim to discuss here: to argue against the *uncritical* development of logical modelings along the classical paths of mathematical logic, presenting them as a contribution to the forging of tools useful for applications—also in presence of uncertainty and approximations—in particular when such approaches are intended to be applied in computational intelligence.

## 3 Rigor and the Legacy of Mathematical Logic

These kinds of investigations can be very interesting and pose very challenging questions. The point we aim to focus, then, has nothing to do either with their legitimacy or with the value of the results obtained in themselves. The point we pose is their (claimed) *useful* role *in applications*. Just to clarify what we have in mind, let us see what Hajek wrote in [8], a contribution to a comprehensive volume (“A Companion to Philosophical Logic”) in which he defends the respectability of Fuzzy logic—from the point of view of the logician—when it is *adequately* interpreted. After writing at the beginning “In spite of several successful applications, the logician may (and should) ask: is this really a logic? Does it have foundations, mathematical and/or philosophical? I shall try to give a positive answer to this question, at least as mathematical foundations are concerned” (p. 595). And he concludes the paper by writing:

"Fuzzy logic in the narrow sense is a logic, a logic with a comparative notion of truth. It is mathematically deep, inspiring and in quick development. [...] The bridge between fuzzy logic in the broad sense and pure symbolic logic is being built and the results are promising." (p. 604).

Let's go back to the beginning of this same paper. Introducing the difference of Fuzzy logic in a *broad and narrow sense*, Hajek writes: "It turned out that one has to distinguish two notions of fuzzy logic. It was again Zadeh who coined the terms 'fuzzy logic in broad (or wide) and narrow sense': In a broad sense, the term 'fuzzy logic' has been used as synonymous with 'fuzzy set theory and its applications', ... in the emerging narrow sense, fuzzy logic is understood as a theory of approximate reasoning based on many-valued logic. Zadeh [...] stresses that the questions of fuzzy logic in the narrow sense differ from usual questions of many-valued logic and concern more questions of approximate inferences than those of completeness, etc.; with full admiration to Zadeh's pioneering and extensive work [...] a logician will first study classical logical questions on completeness, decidability, complexity, etc. of the symbolic calculi in question and then try to reduce the question of Zadeh's agenda to questions of deduction as far as possible" (p. 596). Let us, finally, look at a few other observations borrowed from another—relatively recent—paper by the same Hajek coauthored with Paris and Shepherdson [9]. The authors, all mathematical logicians by trade, arrive at an interesting conclusion when they write: "our results appear to document the fact that fuzzy logic, taken seriously, is not just applied logic but may well be considered a branch of philosophical logic (since it offers a rich formal model of consequence under vagueness) as well as of mathematical logic (since it brings problems demanding non-trivial mathematical solutions)." [9, p. 341] We advance the opinion that "fuzzy logic, taken seriously", that is what is usually called "fuzzy mathematical logic", independently from the importance and value of its results, is not at all an *applied* logic and has, in practice, nothing to do with applications (and, as a consequence, with a general *explicatum*, in Carnap's sense, of the informal notion of *reasoning*). It cannot provide, in fact, any *true* help for applications in situations where uncertainty and fuzziness play a inavoidable role at least for the following two simple reasons: (1) the crucial concepts of mathematical logic, soundness and completeness lose their crucial role (at least for applications) when we are concerned from the start with a pervasive presence of uncertainty and imprecision. (2) Secondly, all this complex (and wonderful) machinery complicates (if taken seriously) any approach to solve any non trivial problem. This fact should be afforded, anyway, if we think that the approach can produce better results, but this is not the case in view of point 1 above (as well as with the experience done). The crucial points of mathematical logic are not motivated by original, general aspects of the informal notion of reasoning. Its agenda—when it was conceived—was different, and was different since it was dictated by the needs of the Hilbert program: *to look for certainty*. There was the need of an important insurance: to avoid unpleasant situations (the paradoxes), when the mathematicians were doing their job of "searching proofs" also in the new territories beyond the frontier opened to them by Cantor. The same can be said also for "fuzzy mathematical logic" which has modelled itself on the same standards of classical mathematical logic. But—from the point of view of applications and the modeling of reasoning in real life situations—the search for a *rigorous* “approximate certainty”—as we could call it—looks peculiar. Let us observe that we have not (and shall not) discuss the problem of probabilistic reasoning. This could seem strange at first sight. We thought about it; however, after a reflection we realized that the point we wanted to focus was the development of models of reasoning based on mathematical logic (when uncertainty and imprecision are *added*). Probabilistic reasoning takes into account from the start the presence of uncertainty, and so the approaches based on probabilistic considerations represent and are a *different* chapter of this story. Of course, along the way of constructing a (general) theory of reasoning these *nuances* should be considered. But, at the moment we want concentrate to this specific aspect of the problem. For clarity reasons we shall, however, in the brief additional comments which follows, indicate a few connections with probabilistic attitudes which naturally emerge.

## 4 Focus on a Few New Crucial Notions

For brevity let us boldly say that in trying to construct a general theory of reasoning in a fresh way there is no need of looking for pivoting notions. In the case of Hilbert the questions of soundness and completeness were essential. In the case of everyday’s (although rigorous) use of reasoning, the situation is different. It can happen, of course, that we meet the emergence of new ideas upon which it is useful to pay attention. But it seems that it is not useful to start from them. At the moment we shall limit ourselves to comment on a few remarks of von Neumann [11]. It can be helpful, in future, to rethink a few considerations done by Bellmann and Zadeh [4] on the notion of “locality”. It is well known that von Neumann in his last years was deeply involved in cybernetic questions and the design of computers, and pondered over logic and the way in which it could be modified for being used in the new emerging fields. In particular he wished for logic a development that could allow the use of methods over the continuum and of mathematical analysis. Now we can say that—in its general lines—this project has been at least partly realized, although we do not know whether the present accomplishments have been done in the direction he had in mind. We refer to the introduction of generalized connectives in fuzzy logic starting from the seminal paper by Trillas on negation [19] (and the subsequent generalization of other connectives [1]).

We must remember that von Neumann also gave thought to another crucial question, the presence of error and the way of treating it. Can we affirm that this has something to do with the questions we are discussing now? Let us consider now an apparently different question. All logical systems have developed themselves with the idea in mind that a particular and careful attention should be paid to the fact that, in a sense, there is a flow of truth when applying correct rules. A truth “transmitted” from axioms, or other intermediate steps to the conclusions searched for. This is used, in Italian, in the title of a remarkable book about logic by Bellissima and Pagli [3], a wordplay on the mainly religious concept of “received truth.” Now, it seems to us that in reasoning with partial information, this beautiful metaphor is no more valid. There is really nothing to be rigourously transmitted, unless we consider not the transmission of truth, but the way in which uncertainty plays its game. In this case we could also use the image of how uncertainty flows from the premises to the conclusions for taking into account the way in which uncertainty impact on our way of drawing conclusions. This is the place in which probability (probabilistic rules and concepts) can play an essential role. The uncertainty present, in fact, must be quantitatively represented and modelled through one of the available theories and approaches which appear to correspond better to the specific situation taken into account [5–7]. One should consider in this context also psychological aspect of the empirical phenomenon of reasoning which can be useful to develop and enrich computational intelligence models [2, 12]. However, the most adequate setting for studying the flow and control of the uncertainty present is one in which one must look to a subtle play of checks and balances between old and new information and the way in which these changes have an impact on some new questions. It is really something that has to do with von Neumann suggestion that we must take error seriously and treat it through thermodynamical methods. In this context it seems interesting to look at the connection between fuzzy sets and subjective probability found by Coletti and Scozzafava (see [5]) as well as to remember that the logical (extended) connective used by Mamdani (see e.g. [10]) in his applications of fuzzy techniques to control theory is more similar to a correlation than a true logical implication.

## 5 As a Sort of (Provisional) Conclusion

The word “conclusion” is not very apt for the sort of considerations and remarks done in the present paper. The comments which follow, then, have only the aim of focusing the small steps forwards have been done (in our view) in order to face the problem with a better awareness. The considerations done in the last Section allow to proceed along the way of asking the crucial questions of Computational intelligence, something we have discussed in the past [14–18, 21], without the burdensome legacy of something which appears to be improper. The challenging question of constructing a unified and unifying approach (and also true models) to intelligent behaviour of both humans and artifacts remains more difficult than it can superficially appear (and appeared at the birth of Cybernetics). Also the possibility of formalizing vagueness in a easy way, as proposed and done by fuzzy sets theory, presents still unanswered questions. However we must get rid of rigidities which do not strictly belong to the main problem. A step forward can be done—we think—if, when trying to model *reasoning*, we look at mathematics and logic in a new way. New means, first of all, in a way similar to the one in which people looked at mathematics and logic before Hilbert programme was started. Limitation theorems have been wonderful intellectual achievements, but for what regards such questions, as the modelling of aspects of intelligent behaviour, they indicate the existence of “boundaries”, not suggestions for grasping essential features of these new domains of Nature we want to understand and model. We need a completely new start, not an adaptation of the technical results of classical mathematical logic. Lotfi Zadeh was right in distinguishing fuzzy logic in the narrow sense and in the wide sense. However, fuzzy logic in the wide sense is not the answer. To recognize that uncertainty, vagueness, fuzziness and imprecision all play an essential role in intelligent behaviour forces us to afford the problem of modeling "reasoning" in a completely fresh way. By starting from the informal idea as used, rigorously with respect to the intended field, in all the contexts of human life and use of natural language [20], one should proceed, through an experimental study, to pick up its meaningful features and proceed along the ways that could help deepening our understanding of this crucial notion.

## References

1. Alsina C, Trillas E, Valverde L (1983) On some logical connectives for fuzzy sets theory. J Math Anal Appl 93(1):15–26
2. Baratgin J, Over DE, Politzer G (2013) Uncertainty and the de finetti tables. Think Reason 19(3–4):308–328
3. Bellissima F, Pagli P (1993) La verità trasmessa. Sansoni
4. Bellman RE, Zadeh LA (1977) Modern uses of multiple-valued logic. In: Local and fuzzy logics. Springer Netherlands, pp 103–165
5. Coletti G, Vantaggi B (2013) Inference with probabilistic and fuzzy information. In: On fuzziness. Springer, pp 115–119
6. Coletti G, Scozzafava R, Vantaggi B (2015) Possibilistic and probabilistic logic under coherence: default reasoning and system p. Mathematica Slovaca 65(4):863–890
7. Gilio A, Pfeifer N, Sanfilippo G (2016) Transitivity in coherence-based probability logic. J Appl Logic 14:46–64
8. Hajek P (2006) Why fuzzy logic? In: Jacquette D (ed) A companion to philosophical logic. Blackwell Publishing Ltd, Oxford, UK
9. Hájek P, Paris J, Shepherdson J (2000) The liar paradox and fuzzy logic. J Symbolic Logic 65:339–346. doi:10.2307/2586541
10. Mamdani A, Gaines B (1981) Fuzzy reasoning and its applications. Academic Press
11. von Neumann J (1956) Probabilistic logics and the synthesis of reliable organisms from unreliable components. Automata Stud 34:43–98
12. Pfeifer N (2013) The new psychology of reasoning: a mental probability logical perspective. Think Reason 19(3–4):329–345
13. Popper K (1993) Unended quest: intellectual autobiography. Routledge
14. Seising R, Tabacchi ME (2013) The webbed emergence of fuzzy sets and computer science education from electrical engineering. In: Ciucci D, Montero J, Pasi G (eds) Proceedings of 8th EuSFLaT. Atlantis Press
15. Seising R, Tabacchi ME, Termini S, Trillas E (2015) Fuzziness, cognition and cybernetics: a historical perspective. In: Alonso J et al (eds) Proceedings of IFSA-EuSFLaT, vol 89. Atlantis Press, AISR, pp 1407–1412
16. Tabacchi ME, Termini S (2015) Experimental modelling for a natural landing of fuzzy sets in new domains. In: Magdalena L, Verdegay JL, Esteva F (eds) Enric trillas: a passion for fuzzy sets, studies in fuzziness and soft computing, vol 322. Springer, pp 179–188
17. Tabacchi ME, Termini S (2015b) Fifty fuzzily gone, many more to go—an appreciation of fuzziness' present and an outlook to what may come. Informatik Spektrum 38(6):484–489. doi:10.1007/s00287-015-0933-6

18. Tabacchi ME, Termini S (2015) Future is where concepts, theories and applications meet (also in fuzzy logic). In: Kacprycz J, Trillas E, Seising R (eds) Towards the future of fuzzy logic, studies in fuzziness and soft computing, vol 325. Springer
19. Trillas E (1998) On negation functions in fuzzy logic. In: Barro S et al (eds) Advances in fuzzy logic, Avances en..., vol 5. Universidade de Santiago de Compostela
20. Trillas E (1998) On negation functions in fuzzy set theory. In: Advances of fuzzy logic, pp 31–43
21. Trillas E, Termini S, Tabacchi ME, Seising R (2015) Fuzziness, cognition and cybernetics: an outlook on future. In: Alonso J et al (eds) Proceedings of IFSA-EuSFLaT, vol 89. Atlantis Press, AISR, pp 1413–1418

# Lexicographic Choice Functions Without Archimedeanicity

**Arthur Van Camp, Enrique Miranda and Gert de Cooman**

**Abstract** We investigate the connection between choice functions and lexicographic probabilities, by means of the convexity axiom considered by Seidenfeld et al. (Synthese 172:157–176, 2010 [7]) but without imposing any Archimedean condition. We show that lexicographic probabilities are related to a particular type of sets of desirable gambles, and investigate the properties of the coherent choice function this induces via maximality. Finally, we show that the convexity axiom is necessary but not sufficient for a coherent choice function to be the infimum of a class of lexicographic ones.

**Keywords** Choice functions · Lexicographic probabilities · Archimedeanicity · Maximality

## 1 Introduction

A prominent decision model under uncertainty is that of *choice functions* [5]. To be able to deal with imprecise information, Seidenfeld et al. proposed an axiomatisation of coherent choice functions in [7] that generalised Rubin's [5] to allow for incomparability. They also established a representation theorem of coherent choice functions by means of probability/utility pairs.

From an imprecise probabilities perspective, choice functions can be seen as a more general model than sets of desirable gambles, because preferences are not uniquely determined by pairwise comparisons between options. We investigated this

A. Van Camp (✉) · G. de Cooman
Ghent University, Data Science Lab, Technologiepark-Zwijnaarde 914,
9052 Zwijnaarde, Belgium
e-mail: Arthur.VanCamp@UGent.be

G. de Cooman
e-mail: Gert.deCooman@UGent.be

E. Miranda
Department of Statistics and Operations Research, University of Oviedo, Oviedo, Spain
e-mail: mirandaenrique@uniovi.es

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_59

idea in [10], and in particular we studied the connections between choice functions and the notions of desirability and indifference. In order to do so, we applied the above-mentioned axiomatisation [7] to gambles instead of horse lotteries, and also removed two axioms: (i) the Archimedean one, because it prevents choice functions from modelling the preferences captured by coherent sets of desirable gambles; and (ii) the convexity axiom, because that is incompatible with maximality as a decision rule, something that is closely tied in with coherent sets of desirable gambles. Although this alternative axiomatisation is more general, it also has the drawback of not leading to a Rubinesque representation theorem, or in other words, to a strong belief structure [2].

In the present paper, we add more detail to our previous findings [10] by investigating in more detail the implications of the convexity axiom, while still letting go of archimedeanicity. We show that, if a Rubinesque representation theorem were possible, it would involve lexicographic probabilities, but that unfortunately such a representation is not generally guaranteed. In establishing this, we derive some properties of coherent choice functions in terms of their so-called rejection sets.

The paper is organised as follows: in Sect. 2, we provide the basics of the theory of choice functions that we need for the rest of the paper. The connection with lexicographic probabilities and the connection with a representation theorem is addressed in Sect. 3. Some additional comments and remarks are provided in Sect. 4. Due to limitations of space, many of the proofs have been omitted.

## 2 Coherent Choice Functions

Consider a finite possibility space $\mathcal{X}$ in which a random variable $X$ takes values. We denote by $\mathcal{L}$ the set of all gambles—real-valued functions—on $\mathcal{X}$. Typically, a gamble $f(X)$ is interpreted as an uncertain reward: if the actual outcome turns out to be $x$ in $\mathcal{X}$, then the subject's capital changes by $f(x)$. For any two gambles $f$ and $g$, we write $f \leq g$ when $f(x) \leq g(x)$ for all $x$ in $\mathcal{X}$, and we write $f < g$ when $f \leq g$ and $f \neq g$. We collect all gambles $f$ for which $f > 0$ in $\mathcal{L}_{>0}$.

For a subset $O$ of $\mathcal{L}$, we define its *positive hull* as $\operatorname{posi}(O) \coloneqq \left\{\sum_{k=1}^{n} \lambda_k f_k : n \in \mathbb{N}, \lambda_k \in \mathbb{R}_{>0}, f_k \in O\right\} \subseteq \mathcal{L}$, and its *convex hull* as $\operatorname{CH}(O) \coloneqq \left\{\sum_{k=1}^{n} \alpha_k f_k : n \in \mathbb{N}, \alpha_k \in \mathbb{R}_{\geq 0}, \sum_{k=1}^{n} \alpha_k = 1, f_k \in O\right\} \subseteq \mathcal{L}$, where $\mathbb{R}_{>0}$ ($\mathbb{R}_{\geq 0}$) is the set of all positive (non-negative) real numbers. For any two subsets $O_1$ and $O_2$ of $\mathcal{L}$ and any $\lambda$ in $\mathbb{R}$, we let $\lambda O_1 \coloneqq \{\lambda f : f \in O_1\}$ and $O_1 + O_2 \coloneqq \{f + g : f \in O_1, g \in O_2\}$.

We denote by $\mathcal{Q}$ the set of all non-empty *finite* subsets of $\mathcal{L}$. Elements $O$ of $\mathcal{Q}$ are the option sets amongst which a subject can choose his preferred options.

**Definition 1** A *choice function* $C$ is a map $C\colon \mathcal{Q} \to \mathcal{Q} \cup \{\emptyset\}\colon O \mapsto C(O)$ such that $C(O) \subseteq O$.

The interpretation is that a choice function $C$ selects the set $C(O)$ of 'best' options in the *option set* $O$. Our definition resembles the one commonly used in the literature [1, 7, 9], except for a (also not unusual) restriction to *finite* option sets [6, 8].

Equivalently to a choice function $C$, we consider its *rejection function* $R$, defined $R(O) := O \setminus C(O)$ for all $O$ in $\mathcal{Q}$. It returns the gambles that are not selected by $C$.

In this paper, we focus on coherent choice functions.

**Definition 2** We call a choice function $C$ on $\mathcal{Q}$ *coherent* if for all $O, O_1, O_2$ in $\mathcal{Q}$, $f, g$ in $\mathcal{L}$ and $\lambda$ in $\mathbb{R}_{>0}$:

C$_1$. $C(O) \neq \emptyset$;

C$_2$. if $f < g$ then $\{g\} = C(\{f, g\})$;

C$_3$. a. if $C(O_2) \subseteq O_2 \setminus O_1$ and $O_1 \subseteq O_2 \subseteq O$ then $C(O) \subseteq O \setminus O_1$;

b. if $C(O_2) \subseteq O_1$ and $O \subseteq O_2 \setminus O_1$ then $C(O_2 \setminus O) \subseteq O_1$;

C$_4$. a. if $O_1 \subseteq C(O_2)$ then $\lambda O_1 \subseteq C(\lambda O_2)$;

b. if $O_1 \subseteq C(O_2)$ then $O_1 + \{f\} \subseteq C(O_2 + \{f\})$.

These axioms are a subset of the ones studied by Seidenfeld et al. [7], translated from horse lotteries to gambles. We have not included the Archimedean axiom, which makes our definition more general. This is important in order to make the connection with the sets of desirable gambles we recall below.

In this paper, we intend to investigate in some detail the implications of an additional axiom in [7], namely

C$_5$. if $O \subseteq O_1 \subseteq \mathrm{CH}(O)$ then $C(O) \subseteq C(O_1)$ for all $O$ and $O_1$ in $\mathcal{Q}$,

also referred to as the *convexity axiom*. One useful property we shall have occasion to use further on is the following:

**Proposition 1** *Let $C$ be a choice function on $\mathcal{L}$ satisfying* C$_3$*a,* C$_4$a *and* C$_5$. *Then for any $n \in \mathbb{N}, f_1, f_2, \ldots, f_n \in \mathcal{L}$ and $\lambda_1, \lambda_2, \ldots \lambda_n \in \mathbb{R}_{>0}$:*

$$0 \in C(\{0, f_1, f_2, \ldots, f_n\}) \Leftrightarrow 0 \in C(\{0, \lambda_1 f_1, \lambda_2 f_2, \ldots, \lambda_n f_n\}).$$

For two choice functions $C$ and $C'$, we call $C$ *not more informative* than $C'$—and we write $C \sqsubseteq C'$—if $C(O) \supseteq C'(O)$ for all $O$ in $\mathcal{Q}$. The binary relation $\sqsubseteq$ is a partial order, and for any collection $\mathcal{C}'$ of choice functions, its infimum $\inf \mathcal{C}'$ exists, and is given by $\inf \mathcal{C}'(O) = \bigcup_{C \in \mathcal{C}'} C(O)$ for all $O$ in $\mathcal{Q}$. Coherence is preserved under arbitrary infima [10, Proposition 3], and it is easy to show that so is convexity:

**Proposition 2** *For any collection $\mathcal{C}'$ of choice functions that satisfy* C$_5$*, its infimum* $\inf \mathcal{C}'$ *satisfies* C$_5$ *as well.*

One important way of defining coherent choice functions is by means of sets of desirable gambles. This connection is explored in some detail in [10]. A set of desirable gambles $D$ is simply a subset of the vector space of gambles $\mathcal{L}$. The underlying idea is that a subject finds every gamble $f$ in her set of desirable gambles strictly better than the status quo—she has a strict preference for the uncertain reward $f$ over 0. As we did for choice functions, we pay special attention to *coherent* sets of desirable gambles, see for instance [3] for a detailed discussion.

**Definition 3** ([3]) A set of desirable gambles $D$ is called *coherent* when $D = \operatorname{posi}(D \cup \mathcal{L}_{>0})$ and $0 \notin D$. We collect all coherent sets of desirable gambles in the set $\bar{\mathcal{D}}$.

We may associate with any $D \in \bar{\mathcal{D}}$ a strict partial order $\prec_D$ on $\mathcal{L}$, by letting $f \prec_D g \Leftrightarrow 0 \prec_D g - f \Leftrightarrow g - f \in D$, so $D = \{f \in \mathcal{L} : 0 \prec_D f\}$; see for instance [3]. This correspondence is one-to-one.

We may also associate with a coherent set of desirable gambles $D$ a choice function $C_D$ based on maximality. For any $O$ in $\mathcal{Q}$, we let $C_D(O)$ be the set of gambles that are undominated, or maximal, in $O$:

$$C_D(O) := \{f \in O : (\forall g \in O) g - f \notin D\} = \{f \in O : (\forall g \in O) f \not\prec_D g\}.$$

Interestingly, the coherent choice function $C_D$ associated with a coherent set of desirable gambles $D$ need not satisfy $\mathrm{C}_5$:

**Proposition 3** *For any coherent set of desirable gambles $D$, its corresponding choice function $C_D$ satisfies* $\mathrm{C}_5$ *if and only* $\operatorname{posi}(D^c) = D^c$.

## 3 Lexicographic Choice Functions

Let $\bar{\mathcal{D}}_{\mathrm{L}} := \{D \in \bar{\mathcal{D}} : \operatorname{posi}(D^c) = D^c\}$. It follows from [4, Proposition 6] that a set of gambles $D \in \bar{\mathcal{D}}_{\mathrm{L}}$ induces a *linear prevision*—an expectation operator with respect to a finitely additive probability—by means of the formula $P_D(f) := \sup\{\mu \in \mathbb{R} : f - \mu \in D\}$ for all $f$ in $\mathcal{L}$. We can make an even tighter connection with the so-called *lexicographic probabilities*.

A *lexicographic probability system* is an $\ell$-tuple $p = (p_1, \dots, p_\ell)$ of probability mass functions on $\mathcal{X}$. We associate with $p$ its expectation operator $\mathrm{E}_p = (\mathrm{E}_{p_1}, \dots, \mathrm{E}_{p_\ell})$, and its preference relation $\prec$ on $\mathcal{L}$:

$$f \prec g \Leftrightarrow \mathrm{E}_p(f) <_{\mathrm{L}} \mathrm{E}_p(g) \text{ for all } f \text{ and } g \text{ in } \mathcal{L},$$

where $<_{\mathrm{L}}$ denotes the usual lexicographic order between $\ell$-tuples.

**Proposition 4** *Given a lexicographic probability system $(p_1, \dots, p_\ell)$, the set of desirable gambles $D := \{f \in \mathcal{L} : 0 \prec f\}$ associated with the preference relation $\prec$ is an element of $\bar{\mathcal{D}}_{\mathrm{L}}$. Conversely, given a set of desirable gambles $D$ in $\bar{\mathcal{D}}_{\mathrm{L}}$, its associated preference relation $\prec_D$ is a preference relation based on some lexicographic probability system.*

Because of this result, we refer to the elements of $\bar{\mathcal{D}}_{\mathrm{L}}$ as *lexicographic sets of desirable gambles*, and call the elements of $\bar{\mathcal{C}}_{\mathrm{L}} := \{C_D : D \in \bar{\mathcal{D}}_{\mathrm{L}}\}$ *lexicographic choice functions*.

We gather from the discussion in Sect. 2 that the infimum of any set of lexicographic choice functions satisfies Axioms $\mathrm{C}_1$–$\mathrm{C}_5$. The central question that remains now, is whether any choice function that satisfies Axioms $\mathrm{C}_1$–$\mathrm{C}_5$ is, conversely, an infimum of lexicographic choice functions. Such a representation result would make lexicographic choice functions fulfil the role of *'dually atomic' choice functions* in our theory without the Archimedean axiom, in analogy with the theory with an Archimedean axiom [7], where the dually atomic choice functions are the ones induced by probability mass functions—see [2] for the terminology. In other words, we study the following:

> Is, in parallel with the result in [7], every choice function $C$ that satisfies Axioms $\mathrm{C}_1$–$\mathrm{C}_5$ an infimum of lexicographic choice functions, or in other words, is $C(O) = \bigcup\{C'(O) : C' \in \bar{\mathcal{C}}_{\mathrm{L}}, C \sqsubseteq C'\}$ for all $O$ in $\mathcal{Q}$?

We now show that this is not the case. In our counterexample, we focus on a binary space $\mathcal{X} = \{a, b\}$. It follows from the axioms of coherence that any coherent choice function $C$ on a binary possibility space $\mathcal{X}$ can be determined by two sets: its associated set of desirable gambles $D_C \coloneqq \{f \in \mathcal{L} : \{f\} = C\{0, f\}\}$ and a so-called *rejection set* $K$, which consists of the gambles $g$ in $\mathcal{L}_{\mathrm{II}}$ and $h$ in $\mathcal{L}_{\mathrm{IV}}$ which, taken alone, do not allow us to reject 0, but taken together, do allow us to reject 0:

$$0 \in C(\{0, g\}), 0 \in C(\{0, h\}), \text{ and } 0 \in R(\{0, g, h\}).$$

Here $\mathcal{L}_{\mathrm{II}} \coloneqq \{h \in \mathcal{L} : h(a) < 0 \text{ and } h(b) > 0\}$ constitutes the second, and $\mathcal{L}_{\mathrm{IV}} \coloneqq \{h \in \mathcal{L} : h(a) > 0 \text{ and } h(b) < 0\}$ the fourth quadrant, in the two-dimensional vector space $\mathcal{L}$.

In order to construct our counterexample, consider some increasing subset $K$ of $\mathbb{R}_{>0} \times \mathbb{R}_{<0}$, and use it to define a special choice function $C_K$, with rejection function $R_K$, as follows. First of all, for any option set $O$, we let $0 \in R_K(\{0\} \cup O)$ if and only if

$$O \cap \mathcal{L}_{>0} \neq \emptyset \text{ or } (\exists \lambda_1, \lambda_2 \in \mathbb{R}_{>0})(\exists (\rho_1, \rho_2) \in K)\{\lambda_1(-1, \rho_1), \lambda_2(1, \rho_2)\} \subseteq O. \tag{1}$$

Of course, this will define a choice function $C_K$ *uniquely*, provided that we require that $C_K$ should satisfy Axiom $\mathrm{C}_4$b, because then, for any $O \in \mathcal{Q}$ and any $f \in O$:

$$f \in R_K(O) \Leftrightarrow 0 \in R_K(\{0\} \cup O'), \tag{2}$$

where $O' \coloneqq (O - \{f\}) \setminus \{0\}$.

**Proposition 5** *Any choice function $C_K$ that is defined by Eqs. (1) and (2) satisfies Axioms* $\mathrm{C}_1$, $\mathrm{C}_2$, $\mathrm{C}_3$a, $\mathrm{C}_4$a *and* $\mathrm{C}_4$b.

As far as $\mathrm{C}_5$ is concerned, we have established the following:

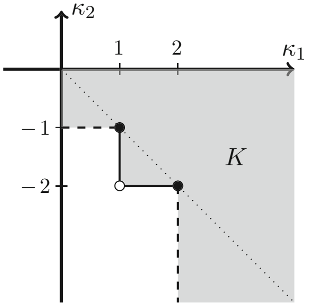


**Fig. 1** The rejection set $K$ that defines the choice function $C_K$ in Proposition 7

**Proposition 6** *Consider any increasing* $K \subseteq \mathbb{R}_{>0} \times \mathbb{R}_{<0}$*. For the choice function* $C_K$ *on* $\mathcal{X} = \{a, b\}$ *defined by Eqs. (1) and (2), the following statements are equivalent:*

*(i)* $C_K$ *satisfies* $\text{C}_5$.
*(ii)* $(\forall(\kappa_1, \kappa_2) \in \mathbb{R}_{>0} \times \mathbb{R}_{<0})(\kappa_1 + \kappa_2 > 0 \Rightarrow (\kappa_1, \kappa_2) \in K)$.

Now, let us consider the set $K$ as depicted in the figure above (Fig. 1).

Let $C_K$ be the choice function associated with this set by means of Eqs. (1) and (2). It follows from the discussion above that this $C_K$ satisfies Axioms $\text{C}_1$, $\text{C}_2$, $\text{C}_3$a, $\text{C}_4$a, $\text{C}_4$b and $\text{C}_5$. Let us show that it also satisfies Axiom $\text{C}_3$b.

**Proposition 7** $C_K$ *satisfies Axiom* $\text{C}_3$b. *As a consequence, it is a coherent choice function that satisfies* $\text{C}_5$.

*Proof* It can be checked that Axiom $\text{C}_3$b is equivalent to

$$(\forall O \in \mathcal{Q}, \forall g \in O)\{0, g\} \subseteq R(O) \Rightarrow 0 \in R(O \setminus \{g\}).$$

So assume that $\{0, g\} \subseteq R_K(O)$. Then $g \in R_K(O)$ and there are $(\kappa_1, \kappa_2) \in K$ such that $\{\lambda_1(-1, \kappa_1), \lambda_2(1, \kappa_2)\} \subseteq O$ for some $\lambda_1$ and $\lambda_2$ in $\mathbb{R}_{>0}$.

If $g \neq \lambda_1(-1, \kappa_1)$ and $g \neq \lambda_2(1, \kappa_2)$ then $0 \in R_K(O \setminus \{g\})$ and we are done, so assume that $g = \lambda_1(-1, \kappa_1)$ or $g = \lambda_2(1, \kappa_2)$.

If $g = \lambda_1(-1, \kappa_1)$, then $0 \in R_K(O - \{g\})$, so there are $(\kappa_1', \kappa_2') \in K$ such that $\{g + \lambda_1'(-1, \kappa_1'), g + \lambda_2'(1, \kappa_2')\} \subseteq O$ for some $\lambda_1'$ and $\lambda_2'$ in $\mathbb{R}_{>0}$, implying that $\{(-\lambda_1 - \lambda_1', \lambda_1\kappa_1 + \lambda_1'\kappa_1'), (-\lambda_1 + \lambda_2', \lambda_1\kappa_1 + \lambda_2'\kappa_2')\} \subseteq O$.

We now have a number of possibilities for the $K$ defined in the figure above.

First of all, $(\frac{\lambda_1\kappa_1 + \lambda_1'\kappa_1'}{\lambda_1 + \lambda_1'}, \kappa_2) \in K$ under any of the following conditions:

(i) $\kappa_2 > -1$;
(ii) $\kappa_2 \in (-2, -1]$ (so $\kappa_1 \geq 1$) and $\kappa_1' \geq 1$;
(iii) $\kappa_2 = -2$ (so $\kappa_1 > 1$) and $\kappa_1' \geq 1$;
(iv) $\kappa_2 < -2$ (so $\kappa_1 > 2$) and $\kappa_1' \geq 2$.

So, in any of these cases, we see that $0 \in R_K(\{(-1, \frac{\lambda_1\kappa_1 + \lambda_1'\kappa_1'}{\lambda_1 + \lambda_1'})\}, 0, (1, \kappa_2))$, and therefore also $0 \in R(\{g + \lambda_1'(-1, \kappa_1'), 0, \lambda_2(1, \kappa_2)\})$, by Proposition 1. Since $\lambda_1'(-1, \kappa_1') \neq 0$, we infer from Axiom $\text{C}_3$a that indeed $0 \in R_K(O \setminus \{g\})$.

The remaining two possibilities are:

(v) $\kappa_2 \leq -1$ (so $\kappa_1 \geq 1$) and $\kappa_1' < 1$ (so $\kappa_2' > -1$);
(vi) $\kappa_2 < -2$ (so $\kappa_1 > 2$) and $\kappa_1' \in [1, 2)$ (so $\kappa_2' \geq -2$).

There are now three possible cases.

If $\lambda_1 = \lambda_2'$, then $\lambda_1\kappa_1 + \lambda_2'\kappa_2' = \lambda_1(\kappa_1 + \kappa_2') > 0$ and therefore also $(-\lambda_1 + \lambda_2', \lambda_1\kappa_1 + \lambda_2'\kappa_2') > 0$, whence $0 \in R_K(\{0, (-\lambda_1 + \lambda_2', \lambda_1\kappa_1 + \lambda_2'\kappa_2')\})$, by Axiom C$_2$.

If $\lambda_1 < \lambda_2'$, then $(\kappa_1', \frac{\lambda_1\kappa_1+\lambda_2'\kappa_2'}{-\lambda_1+\lambda_2'}) \in K$, and therefore also

$$0 \in R_K\left(\left\{(-1, \kappa_1'), 0, (1, \frac{\lambda_1\kappa_1 + \lambda_2'\kappa_2'}{-\lambda_1 + \lambda_2'})\right\}\right)$$

Proposition 1 now guarantees that also

$$0 \in R_K(\{(-\lambda_1', \lambda_1'\kappa_1'), 0, (-\lambda_1 + \lambda_2', \lambda_1\kappa_1 + \lambda_2'\kappa_2')\}).$$

Since $(-\lambda_1', \lambda_1'\kappa_1') \neq g = (-\lambda_1, \lambda_1\kappa_1)$—because $\kappa_1 \geq 1$ and $\kappa_1' < 1$, or $\kappa_1 > 2$ and $\kappa_1' < 2$, we infer from Axiom C$_3$a that $0 \in R_K(O \setminus \{g\})$.

Finally, if $\lambda_1 > \lambda_2'$, then $(\frac{\lambda_1\kappa_1+\lambda_2'\kappa_2'}{\lambda_1-\lambda_2'}, \kappa_2') \in K$, implying that

$$0 \in R_K\left(\left\{(-1, \frac{\lambda_1\kappa_1 + \lambda_2'\kappa_2'}{\lambda_1 - \lambda_2'}), 0, (1, \kappa_2')\right\}\right).$$

Proposition 1 now guarantees that also

$$0 \in R_K(\{(-\lambda_1 + \lambda_2', \lambda_1\kappa_1 + \lambda_2'\kappa_2'), 0, (\lambda_2', \lambda_2'\kappa_2')\}).$$

Since $(-\lambda_1 + \lambda_2', \lambda_1\kappa_1 + \lambda_2'\kappa_2') \neq g = (-\lambda_1, \lambda_1\kappa_1)$, because $\lambda_2' \neq 0$, we infer from Axiom C$_3$a that indeed $0 \in R_K(O \setminus \{g\})$.

The proof of the case that $g = \lambda_2(1, \kappa_2)$ is similar. □

To see that our $C_K$ is not an infimum of lexicographic choice functions, we use the following property:

**Definition 4** Consider a coherent choice function $C$ and its rejection set $K$. Then $C$ is called *weakly Archimedean* if for all $f \in \mathcal{L}_{\mathrm{II}}$ and $g \in \mathcal{L}_{\mathrm{IV}}$ with $\mathrm{posi}(\{f, g\}) \cap \mathcal{L}_{\geq 0} = \emptyset$:

$$(\forall \epsilon \in \mathbb{R}_{>0})(0 \in R(\{f + \epsilon, 0, g\}) \cap R(\{f, 0, g + \epsilon\})) \Rightarrow 0 \in R(\{f, 0, g\}).$$

We use this name because the property is a strictly weaker version of the Archimedean condition in [7, Axioms 3a and 3b]; it still fulfils the role of a continuity condition, but is weak enough to be still compatible with desirability, a non-Archimedean strict preference.

**Proposition 8** *An infimum of a non-empty set of lexicographic choice functions is weakly Archimedean.*

We now see that our choice function $C_K$ from Proposition 7 is not an infimum of lexicographic choice functions, because it is not weakly Archimedean: note that $\{(1+\epsilon, -2), (1, -2+\epsilon)\} \subseteq K$ for all $\epsilon > 0$, while $(1, -2) \notin K$.

## 4 Discussion

We have studied to which extent it is possible to have a theory of coherent choice functions that (i) as a special case allows for choosing the maximal options in the strict binary preference expressed by the notion of desirability in imprecise probabilities—meaning that we must remove the Archimedean axiom, and that (ii) includes lexicographic probability systems as its basic building blocks. We have shown that such a theory can perfectly well incorporate the convexity axiom from [7], but that this additional axiom is not strong enough to warrant a representation theorem where every choice function is an infimum of lexicographic ones. It is still an open problem to uncover additional axioms that will guarantee such representation. We suspect that our weak archimedeanicity will play an important role in solving it.

**Acknowledgments** The research reported in this paper has been supported by project TIN2014-59543-P.

## References

1. Aizerman M (1984) New problems in the general choice theory. Soc Ch Welf 2:235–282
2. De Cooman G (2005) Belief models: an order-theoretic investigation. Ann Math Art Intell 45:5–34
3. De Cooman G, Quaeghebeur E (2012) Exchangeability and sets of desirable gambles. Int J App Reason 53:363–395
4. Miranda E, Zaffalon M (2010) Notes on desirability and coherent lower previsions. Ann Math Art Intell 60:251–309
5. Rubin H (1987) A weak system of axioms for "rational" behavior and the nonseparability of utility from prior. Stat Risk Model 5:47–58
6. Schwartz T (1972) Rationality and the myth of the maximum. Noûs 6:97–117
7. Seidenfeld T, Schervisch M, Kadane J (2010) Coherent choice functions under uncertainty. Synthese 172:157–176
8. Sen A (1971) Choice functions and revealed preference. Rev Econ Stud 38:307–317
9. Sen A (1977) Social choice theory: a re-examination. Econometrica 45:53–89
10. Van Camp A, De Cooman G, Miranda E, Quaeghebeur E (2015) Modelling indifference with choice functions. In: Proceedings of ISIPTA'15, pp 305–314

# Composition Operator for Credal Sets Reconsidered

**Jiřina Vejnarová**

**Abstract** This paper is the second attempt to introduce the composition operator, already known from probability, possibility, evidence and valuation-based systems theories, also for credal sets. We try to avoid the discontinuity which was present in the original definition, but simultaneously to keep all the properties enabling us to design compositional models in a way analogous to those in the above-mentioned theories. These compositional models are aimed to be an alternative to Graphical Markov Models. Theoretical results achieved in this paper are illustrated by an example.

## 1 Introduction

In the second half of 1990s a new approach to efficient representation of multidimensional probability distributions was introduced with the aim to be alternative to Graphical Markov Modeling. This approach is based on a simple idea: a multidimensional distribution is *composed* from a system of low-dimensional distributions by repetitive application of a special composition operator, which is also the reason why such models are called *compositional models*.

Later, these compositional models were introduced also in possibility theory [7, 8] (here the models are parameterized by a continuous $t$-norm) and almost ten years ago also in evidence theory [3, 4]. In all these frameworks the original idea is kept, but there exist some slight differences among these frameworks.

In [9] we introduced a composition operator for credal sets, but due to the problem of discontinuity it needed a revision. After a thorough reconsideration we decided to present a new proposal avoiding this discontinuity. The goal of this paper is to show that the revised composition operator keeps the basic properties of its counterparts in other frameworks, and therefore it will enable us to introduce compositional models for multidimensional credal sets.

J. Vejnarová (✉)
Institute of Information Theory and Automation of the Czech Academy of Sciences,
Pod Vodárenskou věží 4, Prague, Czech Republic
e-mail: vejnar@utia.cas.cz

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_60

This contribution is organized as follows. In Sect. 2 we summarise the basic concepts and notation. The new definition of the operator of composition is presented in Sect. 3, which is devoted also to its basic properties and an illustrative example.

## 2 Basic Concepts and Notation

In this section we will briefly recall basic concepts and notation necessary for understanding the contribution.

### 2.1 Variables and Distributions

For an index set $N = \{1, 2, \dots, n\}$ let $\{X_i\}_{i \in N}$ be a system of variables, each $X_i$ having its values in a finite set $\mathbf{X}_i$ and $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \cdots \times \mathbf{X}_n$ be the Cartesian product of these sets.

In this paper we will deal with groups of variables on its subspaces. Let us note that $X_K$ will denote a group of variables $\{X_i\}_{i \in K}$ with values in $\mathbf{X}_K = \times_{i \in K} \mathbf{X}_i$ throughout the paper.

Any group of variables $X_K$ can be described by a *probability distribution* (sometimes also called *probability function*)

$$P : \mathbf{X}_K \longrightarrow [0, 1],$$

such that

$$\sum_{x_K \in \mathbf{X}_K} P(x_K) = 1.$$

Having two probability distributions $P_1$ and $P_2$ of $X_K$ we say that $P_1$ is *absolutely continuous* with respect to $P_2$ (and denote $P_1 \ll P_2$) if for any $x_K \in \mathbf{X}_K$

$$P_2(x_K) = 0 \Longrightarrow P_1(x_K) = 0.$$

This concept plays an important role in the definition of the composition operator.

### 2.2 Credal Sets

A *credal set* $\mathcal{M}(X_K)$ describing a group of variables $X_K$ is usually defined as a closed convex set of probability measures describing the values of this variable. In order to simplify the expression of operations with credal sets, it is often considered [5] that a credal set is the set of probability distributions associated to the probability measures in it. Under such consideration a credal set can be expressed as a *convex hull* (denoted by CH) of its extreme distributions (ext)

$$\mathcal{M}(X_K) = \mathrm{CH}\{ext(\mathcal{M}(X_K))\}.$$

Consider a credal $\mathcal{M}(X_K)$. For each $L \subset K$ its *marginal credal set* $\mathcal{M}(X_L)$ is obtained by element-wise marginalization, i.e.

$$\mathcal{M}(X_L) = \mathrm{CH}\{P^{\downarrow L} : P \in \mathrm{ext}(\mathcal{M}(X_K))\}, \tag{1}$$

where $P^{\downarrow L}$ denotes the marginal distribution of $P$ on $\mathbf{X}_L$.

Besides marginalization we will also need the opposite operation, called vacuous extension. *Vacuous extension* of a credal set $\mathcal{M}(X_L)$ describing $X_L$ to a credal set $\mathcal{M}(X_K) = \mathcal{M}(X_L)^{\uparrow K}$ $(L \subset K)$ is the maximal credal set describing $X_K$ such that $\mathcal{M}(X_K)^{\downarrow L} = \mathcal{M}(X_L)$.

Having two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing $X_K$ and $X_L$, respectively (assuming that $K, L \subseteq N$), we say that these credal sets are *projective* if their marginals describing common variables coincide, i.e. if

$$\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L}).$$

Let us note that if $K$ and $L$ are disjoint, then $\mathcal{M}_1$ and $\mathcal{M}_2$ are always projective, as $\mathcal{M}_1(X_\emptyset) = \mathcal{M}_2(X_\emptyset) \equiv 1$.

## 2.3 Strong Independence

Among the numerous definitions of independence for credal sets [1] we have chosen strong independence, as it seems to be the most appropriate for multidimensional models.

We say that (groups of) variables $X_K$ and $X_L$ ($K$ and $L$ disjoint) are *strongly independent* with respect to $\mathcal{M}(X_{K\cup L})$ iff (in terms of probability distributions)

$$\mathcal{M}(X_{K\cup L}) = \mathrm{CH}\{P_1 \cdot P_2 : P_1 \in \mathcal{M}(X_K), P_2 \in \mathcal{M}(X_L)\}.$$

Again, there exist several generalizations of this notion to conditional independence, see e.g. [5], but as the following definition is suggested by the authors as the most appropriate for the marginal problem, it seems to be a suitable concept also in our case, since the operator of composition can also be used as a tool for solution of a marginal problem, as shown (in the framework of possibility theory) e.g. in [8].

Given three groups of variables $X_K, X_L$ and $X_M$ ($K, L, M$ be mutually disjoint subsets of $N$, such that $K$ and $L$ are nonempty), we say that $X_K$ and $X_L$ are *conditionally independent* given $X_M$ under global set $\mathcal{M}(X_{K\cup L\cup M})$ (to simplify the notation we will denote this relationship by $K \perp\!\!\!\perp L|M$) iff

$$\mathcal{M}(X_{K\cup L\cup M})$$
$$= \mathrm{CH}\{(P_1 \cdot P_2)/P_1^{\downarrow M} : P_1 \in \mathcal{M}(X_{K\cup M}), P_2 \in \mathcal{M}(X_{L\cup M}), P_1^{\downarrow M} = P_2^{\downarrow M}\}\,.$$

This definition is a generalisation of stochastic conditional independence: if $\mathcal{M}(X_{K\cup L\cup M})$ is a singleton, then $\mathcal{M}(X_{K\cup M})$ and $\mathcal{M}(X_{L\cup M})$ are also (projective) singletons and the definition reduces to the definition of stochastic conditional independence.

# 3 Composition Operator

In this section we will introduce a new definition of composition operator for credal sets. The concept of the composition operator is presented first in a precise probability framework, as it seems to be useful for better understanding to the concept.

## *3.1 Composition Operator of Probability Distributions*

Now, let us recall the definition of composition of two probability distributions [2]. Consider two index sets $K, L \subset N$. We do not put any restrictions on $K$ and $L$; they may be but need not be disjoint, and one may be a subset of the other. Let $P_1$ and $P_2$ be two probability distributions of (groups of) variables $X_K$ and $X_L$; then

$$(P_1 \triangleright P_2)(X_{K\cup L}) = \frac{P_1(X_K) \cdot P_2(X_L)}{P_2(X_{K\cap L})}, \tag{2}$$

whenever $P_1(X_{K\cap L}) \ll P_2(X_{K\cap L})$; otherwise, it remains undefined.

It is specific property of composition operator for probability distributions—in other settings the operator is always defined [3, 8].

## *3.2 Definition and Example*

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be credal sets describing $X_K$ and $X_L$, respectively. Our goal is to define a new credal set, denoted by $\mathcal{M}_1 \triangleright \mathcal{M}_2$, which will be describing $X_{K\cup L}$ and will contain all of the information contained in $\mathcal{M}_1$ and, as much as possible, in $\mathcal{M}_2$.

The required properties are met by Definition 1 in [9].[1] However, that definition exhibits a kind of discontinuity and was thoroughly reconsidered. Here we decided to propose the following one.

[1] Let us note that the definition is based on Moral's concept of conditional independence with relaxing convexity.

**Definition 1** For two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing $X_K$ and $X_L$, their *composition* $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is defined as a convex hull of probability distributions $P$ obtained as follows. For each couple of distributions $P_1 \in \mathcal{M}_1(X_K)$ and $P_2 \in \mathcal{M}_2(X_L)$ such that $P_2^{\downarrow K\cap L} \in argmin\{Q_2 \in \mathcal{M}_2(X_{K\cap L}) : d(Q_2, P_1^{\downarrow K\cap L})$, distribution $P$ is obtained by one of the following rules:

[a] if $P_1^{\downarrow K\cap L} \ll P_2^{\downarrow K\cap L}$

$$P = \frac{P_1 \cdot P_2}{P_2^{\downarrow K\cap L}},$$

[b] otherwise

$$P \in \text{ext}\{P_1^{\uparrow K\cup L}\}.$$

Function $d$ used in the definition is a suitable distance function (e.g. Kullback-Leibler divergence, total variation or some other f-divergence [6]).

Let us note, that this definition of composition operator does not differ from the original one [9] in case of projective credal sets, as in this case the only distributions in $\mathcal{M}_1 \triangleright \mathcal{M}_2$ are those satisfying $P = (P_1 \cdot P_2)/P_2^{\downarrow K\cap L}$, where $P_1^{\downarrow K\cap L} = P_2^{\downarrow K\cap L}$. However, it differs in the remaining cases. Let us illustrate the application of the operator in case [a] by an example.

*Example 1* Let

$$\mathcal{M}_1(X_1X_2) = \text{CH}\{[0.2, 0.8, 0, 0], [0.1, 0.4, 0.1, 0.4], [0.25, 0.25, 0.25, 0.25], [0, 0, 0.5, 0.5]\},$$

and

$$\mathcal{M}_2(X_2X_3) = \text{CH}\{[0, 0.3, 0, 0.7], [0.2, 0.1, 0.4, 0.3], [0.5, 0, 0.5, 0], [0.2, 0.3, 0.2, 0.3]\},$$

be two credal sets describing binary variables $X_1X_2$ and $X_2X_3$, respectively. These two credal sets are not projective, as $\mathcal{M}_1(X_2) = \text{CH}\{[0.2, 0.8], [0.5, 0.5]\}$, while $\mathcal{M}_2(X_2) = \text{CH}\{[0.3, 0.7], [0.5, 0.5]\}$. Therefore $\mathcal{M}_2(X_2) \subset \mathcal{M}_1(X_2)$. Definition 1 in this case leads (using total variation) to

$$\begin{aligned}
&(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3)\\
&= \text{CH}\{[0, 0.3, 0, 0.7, 0, 0, 0, 0], [0.2, 0.1, 0.4, 0.3, 0, 0, 0, 0],\\
&\quad [0, 0.1, 0, 0.3, 0, 0.2, 0, 0.4], [0.07, 0.03, 0.17, 0.13, 0.13, 0.07, 0.23, 0.17],\\
&\quad [0.25, 0, 0.25, 0, 0.25, 0, 0.25, 0], [0.1, 0.15, 0.1, 0.15, 0.1, 0.15, 0.1, 0.15],\\
&\quad [0, 0, 0, 0, 0.5, 0, 0.5, 0], [0, 0, 0, 0, 0.2, 0.3, 0.2, 0.3]\\
&\quad [0, 0.2, 0, 0.8, 0, 0, 0, 0], [0.13, 0.07, 0.46, 0.34, 0, 0, 0, 0],\\
&\quad [0, 0.1, 0, 0.4, 0, 0.1, 0, 0.4], [0.07, 0.03, 0.23, 0.17, 0.07, 0.03, 0.23, 0.17]\}.
\end{aligned}$$

On the other hand

$$\begin{aligned}&(\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_1X_2X_3)\\ &\quad = \mathrm{CH}\{[0, 0.3, 0, 0.7, 0, 0, 0, 0], [0.2, 0.1, 0.4, 0.3, 0, 0, 0, 0],\\ &\qquad [0, 0.1, 0, 0.3, 0, 0.2, 0, 0.4], [0.07, 0.03, 0.17, 0.13, 0.13, 0.07, 0.23, 0.17],\\ &\qquad [0.25, 0, 0.25, 0, 0.25, 0, 0.25, 0], [0.1, 0.15, 0.1, 0.15, 0.1, 0.15, 0.1, 0.15],\\ &\qquad [0, 0, 0, 0, 0.5, 0, 0.5, 0], [0, 0, 0, 0, 0.2, 0.3, 0.2, 0.3]\},\end{aligned}$$

which differs from $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3)$. ◇

This difference deserves an explanation. $\mathcal{M}_2 \triangleright \mathcal{M}_1$ is smaller (more precise) than $\mathcal{M}_1 \triangleright \mathcal{M}_2$, which corresponds to the idea that we want $\mathcal{M}_2 \triangleright \mathcal{M}_1$ to keep all the information contained in $\mathcal{M}_2$. Therefore, we do not consider those distributions from $\mathcal{M}_1$ not corresponding to any from $\mathcal{M}_2$, although these distributions are taken into account when composing $\mathcal{M}_1 \triangleright \mathcal{M}_2$.

This is an example of a typical property of the operator of composition—it is not commutative. The next subsection is devoted to other basic properties.

### 3.3 Basic Properties

In the following lemma we prove that this composition operator possesses basic properties required above.

**Lemma 1** *For two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing $X_K$ and $X_L$, respectively, the following properties hold true:*

1. *$\mathcal{M}_1 \triangleright \mathcal{M}_2$ is a credal set describing $X_{K\cup L}$.*
2. *$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_K) = \mathcal{M}_1(X_K)$.*
3. *$\mathcal{M}_1 \triangleright \mathcal{M}_2 = \mathcal{M}_2 \triangleright \mathcal{M}_1$ iff $\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L})$.*

*Proof* 1. To prove that $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is a credal set describing $X_{K\cup L}$ it is enough to take into consideration that it is the convex hull of probability distributions on $\mathbf{X}_{K\cup L}$, which is obvious from both [a] and [b] of Definition 1.

2. As marginalization of a credal set is element-wise, it is enough to prove that for any $P \in (\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L})$, $P^{\downarrow K} = P_1 \in \mathcal{M}_1(X_K)$ holds. But it immediately follows in case [a] from the results obtained for precise probabilities (see e.g. [2]). In case [b] it is obvious, as any $P$ belongs to a vacuous extension of $P_1 \in \mathcal{M}_1(X_K)$ to $X_{K\cup L}$.
3. First, let us assume that

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) = (\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cup L}).$$

Then also its marginals must be identical, particularly

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cap L}) = (\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cap L}).$$

Taking into account 2. of this lemma we have

$$\begin{aligned}(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cap L}) &= \left(((\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}))^{\downarrow K}\right)^{\downarrow K\cap L} \\ &= ((\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_K))^{\downarrow K\cap L} \\ &= (\mathcal{M}_1(X_K))^{\downarrow K\cap L} = \mathcal{M}_1(X_{K\cap L})\end{aligned}$$

and similarly

$$(\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L}),$$

from which the desired equality immediately follows.

Let, on the other hand, $\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L})$. In this case only [a] of Definition 1 is applied and for any distribution $P$ of $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L})$ there exist $P_1 \in \mathcal{M}_1(X_K)$ and $P_2 \in \mathcal{M}_2(X_L)$ such that $P_1^{\downarrow K\cap L} = P_2^{\downarrow K\cap L}$ and $P = (P_1 \cdot P_2)/P_2^{\downarrow K\cap L}$. But simultaneously (due to projectivity) $P = (P_1 \cdot P_2)/P_1^{\downarrow K\cap L}$, which is an element of $(\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cup L})$. Hence

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) = (\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cup L}),$$

as desired. □

The following theorem, proven in [9], expresses the relationship between strong independence and the operator of composition. It is, together with Lemma 1, the most important assertion enabling us to introduce multidimensional models.

**Theorem 1** *Let $\mathcal{M}$ be a credal set describing $X_{K\cup L}$ with marginals $\mathcal{M}(X_K)$ and $\mathcal{M}(X_L)$. Then*

$$\mathcal{M}(X_{K\cup L}) = (\mathcal{M}^{\downarrow K} \triangleright \mathcal{M}^{\downarrow L})(X_{K\cup L})$$

*iff*

$$(K \setminus L) \perp\!\!\!\perp (L \setminus K)|(K \cap L).$$

This theorem remains valid also for this, revised definition, as $\mathcal{M}(X_K)$ and $\mathcal{M}(X_L)$ are marginals of $\mathcal{M}(X_{K\cup L})$, and therefore only [a] (for projective distributions) is applicable.

## 4 Conclusions

We presented revised version of composition operator for credal sets. This definition seems to be satisfactory from the theoretical point of view; it satisfies the basic required properties and, in contrary to the original one, it avoids discontinuity.

It seems to be a reasonable tool for construction of compositional multidimensional models. Nevertheless, many problems should be solved in the near future. From the theoretical point of view it is the relationship to probabilistic and evidential compositions operators. From the practical viewpoint it is the problem of effective finding of the nearest probability distributions (if there is no projective).

**Acknowledgments** The support of Grant GAČR 16-12010S is gratefully acknowledged.

## References

1. Couso I, Moral S, Walley P (1999) Examples of independence for imprecise probabilities. In: De Cooman G, Cozman FG, Moral S, Walley P (eds) Proceedings of ISIPTA'99. Ghent, pp 121–130
2. Jiroušek R (1997) Composition of probability measures on finite spaces. In: Geiger D, Shenoy PP (eds) Proceedings of UAI'97. Morgan Kaufmann Publishers, San Francisco, California, pp 274–281
3. Jiroušek R, Vejnarová J, Daniel M (2007) Compositional models for belief functions. In: De Cooman G, Vejnarová J, Zaffalon M (eds) Proceedings of ISIPTA'07. Mat-fyz Press, Praha, pp 243–252
4. Jiroušek R, Vejnarová J (2011) Compositional models and conditional independence in evidence theory. Int J Approx Reason 52:316–334
5. Moral S, Cano A (2002) Strong conditional independence for credal sets. Ann Math Artif Intell 35:295–321
6. Vajda I (1989) Theory of statistical inference and information. Kluwer Academic Publishers, Dordrecht
7. Vejnarová J (1998) Composition of possibility measures on finite spaces: preliminary results. In: Bouchon-Meunier B, Yager RR (eds) Proceedings of IPMU'98. Editions E.D.K. Paris, pp 25–30
8. Vejnarová J (2007) On possibilistic marginal problem. Kybernetika 43(5):657–674
9. Vejnarová J (2013) Operator of composition for credal sets. In: Cozman FG, Denoeux T, Destercke S, Seidenfeld T (eds) Proceedings of ISIPTA'13, pp 355–364

# A Nonparametric Linearity Test for a Multiple Regression Model with Fuzzy Data

**Dabuxilatu Wang**

**Abstract** A linearity test for a multiple regression model with *LR*-fuzzy responses and *LR*-fuzzy explanatory variables is considered. The regression model consists of several multiple regression models from response center or spreads to the explanatory centers and spreads. A multiple nonparametric regression model to be employed as a reference in the testing approach is estimated, and with which the linearity of the regression model is tested. Some simulation example is also presented.

**Keywords** *LR*-fuzzy random variables · Nonparametric regression model · Linearity test

## 1 Introduction

In investigating the relationship between random elements, regression analysis enables to seek for some complex effect of several random elements upon another. Regression techniques have long been relevant to many fields [1, 7]. The random elements considered actually in many practical applications in public health, medical science, ecology, social or economic and financial problems sometimes involve vagueness, so the regression problems have to face with such a mixture of fuzziness and randomness. Under the least squares methods a bivariate linear regression model [8] with $n$-dimensional fuzzy random sets has been estimated, and a multiple linear regression model with *LR*-response variable and crisp explanatory variables or with both *LR*-fuzzy response and explanatory variables is proposed in [4, 5], respectively. In [6] a linearity test for a simple linear model with both interval-valued input and output has been given. In [3] a linearity test for a simple linear model with a *LR*-response variable and a crisp explanatory variable was proposed under a nonparametric method. However, such a linearity test has not been applied for a multiple regression model with fuzzy data [5].

D. Wang (✉)
Department of Statistics, Guangzhou University, No. 230 Waihuanxi Road,
Higher Education Mega Center, Guangzhou 510006, China
e-mail: wangdabu@gzhu.edu.cn

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_61

In this paper, we focus on a nonparametric linearity test for a multiple linear regression model with *LR*-fuzzy number-valued inputs and outputs.

## 2 Preliminaries

Let $\mathbb{R}$ be the set of all real numbers. A fuzzy set on $\mathbb{R}$ is defined to be a mapping $u : \mathbb{R} \to [0, 1]$ satisfying following conditions:

(1) $u_\alpha = \{x|u(x) \geq \alpha\}$ is a closed bounded interval for each $\alpha \in (0, 1]$, i.e. $u_\alpha = [\inf u_\alpha, \sup u_\alpha]$.
(2) $u_0 = supp$ u is a closed bounded interval.
(3) $u_1 = \{x|u(x) = 1\}$ is nonempty.

where *supp* u $= cl\{x|u(x) > 0\}$, *cl* denotes the *closure* of a set. Such a fuzzy set is also called a *fuzzy number*. By $\mathcal{F}(\mathbb{R})$ we denote the set of all fuzzy numbers, with Zadeh's extension principle the arithmetic operation $*$ on $\mathcal{F}(\mathbb{R})$ can be defined by $(u * v)(t) = \sup_{\{t_1,t_2:t_1*t_2=t\}}\{\min(u(t_1), v(t_2))\}$, $u, v \in \mathcal{F}(\mathbb{R}), t, t_1, t_2 \in \mathbb{R}, * \in \{\oplus, \ominus, \odot\}$, where $\oplus, \ominus, \odot$ denote the addition, subtraction and scalar multiplication among fuzzy numbers, respectively.

The following parametric class of fuzzy numbers, the so-called *LR*-fuzzy numbers, are often used in applications:

$$u(x) = \begin{cases} L(\frac{m-x}{l}), & x \leq m \\ R(\frac{x-m}{r}), & x > m \end{cases}$$

Here $L : \mathbb{R}^+ \to [0, 1]$ and $R : \mathbb{R}^+ \to [0, 1]$ are given left- continuous and non-increasing function with $L(0) = R(0) = 1$. $L$ and $R$ are called left and right shape functions, $m$ the central point of $u$ and $l > 0, r > 0$ are the left and right spread of $u$. An *LR*-fuzzy number is abbreviated by $u = (m, l, r)_{LR}$. An *LR*-fuzzy number is said to be symmetric if $L(x) = R(x)$ and $l = r$. It has been proven that:

$$(m_1, l_1, r_1)_{LR} \oplus (m_2, l_2, r_2)_{LR} = (m_1 + m_2, l_1 + l_2, r_1 + r_2)_{LR}$$

$$a \odot (m, l, r)_{LR} = \begin{cases} (am, al, ar)_{LR}, & a > 0 \\ (am, -ar, -al)_{RL}, & a < 0 \\ (0, 0, 0), & a = 0 \end{cases}$$

Let $L(\alpha) := \sup\{x \in \mathbb{R}|L(x) \geq \alpha\}, R(\alpha) := \sup\{x \in \mathbb{R}|R(x) \geq \alpha\}$. Then for $u = (m, l, r)_{LR}, u_\alpha = [m - lL(\alpha), m + rR(\alpha)], \alpha \in [0, 1]$. An *LR-fuzzy random variable* [7] on the probability space $(\Omega, \mathcal{A}, P)$ is defined as a measurable mapping $X : \Omega \to \mathcal{F}_{LR}(\mathbb{R})$, $X(\omega) = (x^m(\omega), x^l(\omega), x^r(\omega))_{LR}$, $\omega \in \Omega$, in brief we denote $X$ as $X = (x^m, x^l, x^r)_{LR}$, where $x^m, x^l, x^r$ are three real-valued random variables with $P\{x^l \geq 0\} = P\{x^r \geq 0\} = 1$.

We will employ the distance between fuzzy numbers $u$ and $v$ proposed by [1, 7] by the $L_2$ metric $\delta_2$,

$$\delta_2(u, v) := \Big(\int_0^1 \int_{\mathbb{S}^0} (S_{u_\alpha}(x) - S_{v_\alpha}(x))^2 \mu(dx) d\alpha\Big)^{1/2},$$

where $\mu$ is a normalized Lebesgue measure, $\mathbb{S}^0 = \{-1, 1\}$, $S_{u_\alpha}(x)$ denotes the support function of $u$. For $u_i = (m_i, l_i, r_i)_{LR}$, $i = 1, 2$, $\delta_2^2(u_1, u_2)=(m_1 - m_2)^2 + \frac{1}{2}L_2(l_1 - l_2)^2 + \frac{1}{2}R_2(r_1 - r_2)^2 - L_1(m_1 - m_2)(l_1 - l_2) + R_1(m_1 - m_2)(r_1 - r_2)$, where $L_2 = \int_0^1 L^2(\alpha)d\alpha, R_2 = \int_0^1 R^2(\alpha)d\alpha, L_1 = \int_0^1 L(\alpha)d\alpha, R_1 = \int_0^1 R(\alpha)d\alpha$. And for the symmetric $u_i = (m_i, l_i)_L$, $i = 1, 2$, $\delta_2^2(u_1, u_2) = (m_1 - m_2)^2 + L_2(l_1 - l_2)^2$.
For *LR*-f.r.v.'s $X = (x^m, x^l, x^r)_{LR}$, $Y = (y^m, y^l, y^r)_{LR}$ the expectation and variance, covariance are defined as follows, $E(X) := \big(E(x^m), E(x^l), E(x^r)\big)_{LR}$, $Var(X) := E(\delta_2^2(X, E(X))) = Varx^m + \frac{1}{2}L_2 Varx^l + \frac{1}{2}R_2 Varx^r - L_1 Cov(x^m, x^l) + R_1 Cov(x^m, x^r)$, $Cov(X, Y) := \int_0^1 (Cov(\inf X_\alpha, \inf Y_\alpha) + Cov(\sup X_\alpha, \sup Y_\alpha))d\alpha = L_2 Cov(x^l, y^l) - L_1(Cov(x^l, y^m) + Cov(x^m, y^l)) + 2Cov(x^m, y^m) + R_1(Cov(x^m, y^r) + Cov(x^r, y^m)) + R_2 Cov(x^r, y^r)$.

## 3 The Nonparametric Linearity Test for the Multiple Regression Model with *LR*-Fuzzy Data [5]

In [8] a bivariate linear regression model with $n$-dimensional fuzzy random sets is estimated. However, the linearity test for this model has not been considered. It is difficult to consider such a test from the estimated model intuitively. We may consider one dimensional case where the fuzzy random variable (f.r.v.) is with a parametric form, the *LR*-f.r.v., and the bivariate consideration can also be extended to the multiple case. Let $\tilde{Y} = (Y^m, Y^l, Y^r)_{LR}$ be a response *LR*-f.r.v., $\tilde{X}_1 = (X_1^m, X_1^l, X^r)_{LR}, \cdots$ $\tilde{X}_p = (X_p^m, X_p^l, X_p^r)_{LR}$ be $p$ explanatory *LR*-f.r.v.'s. On which we have observations $\{\tilde{Y}_i, \tilde{X}_{1i}, \tilde{X}_{2i}, \ldots, \tilde{X}_{pi}\}$, $i = 1, \ldots, n$. A linear relationship between $\tilde{Y}$ and $\tilde{X}_1, \ldots, \tilde{X}_p$ has been approximated by a multiple linear regression model between the response center or response spreads and the explanatory centers and spreads in [5], i.e. the model

$$\begin{cases} Y^m = \underline{Xa}_m^{'} + b_m + \varepsilon^m, \\ g(Y^l) = \underline{Xa}_l^{'} + b_1 + \varepsilon^l, \\ h(Y^r) = \underline{Xa}_r^{'} + b_r + \varepsilon^r, \end{cases} \tag{1}$$

under the condition that the shape functions $L, R$ are predetermined as fixed functions. Where $g, h$ are two invertible functions $g : (0, +\infty) \to \mathbb{R}$ and $h : (0, +\infty) \to \mathbb{R}$, which can be used for transformation of the concerned spread variables [4, 5]. $\underline{X} = (X_1^m, X_1^l, X_1^r, \cdots, X_p^m, X_p^l, X_p^r)$ is the vector of length $3p$ of all the components of the explanatory variables, $\underline{a}_m = (a_{mm}^1, a_{ml}^1, a_{mr}^1, \cdots, a_{mm}^p, a_{ml}^p, a_{mr}^p)$, $\underline{a}_l =$

$(a_{lm}^1, a_{ll}^1, a_{lr}^1, \cdots, a_{lm}^p, a_{ll}^p, a_{lr}^p)$, $\underline{a}_r = (a_{rm}^1, a_{rl}^1, a_{rr}^1, \cdots, a_{rm}^p, a_{rl}^p, a_{rr}^p)$ are vectors of length $3p$ of the unknown parameters related to $\underline{X}$. $\varepsilon^m, \varepsilon^l, \varepsilon^r$ are real valued random variables with $E(\varepsilon^m|\underline{X}) = E(\varepsilon^l|\underline{X}) = E(\varepsilon^r|\underline{X}) = 0$.

On the other hand, the above relationship may be allowed to be in a nonparametric model as follows,

$$\begin{cases} Y^m = f_m(\underline{X}) + \varepsilon^m, \\ g(Y^l) = f_l(\underline{X}) + \varepsilon^l, \\ h(Y^r) = f_r(\underline{X}) + \varepsilon^r, \end{cases} \tag{2}$$

where $\varepsilon^m, \varepsilon^l, \varepsilon^r$ are real valued random variables with mean 0 and variance $\sigma^2$. For the parameters estimation of the models in (1) we refer to [5] under the metric $\delta_2$.

Concerning model in (2), the functions $f_m, f_l, f_r$ could be estimated by means of nonparametric smoothing. In the $s$-variate nonparametric regression model $Y_i = F(t_{1i}, \cdots, t_{si}) + \varepsilon_i$, $i = 1, \cdots, n$, $F$ is a $s$-variate real valued function to be estimated, $\varepsilon_i$ are i.i.d. with mean 0, variance $\sigma^2$, $(t_{1i}, \cdots, t_{si})$ has a density function $f$ with a support set $A$. Considering kernel product, $b_{t_1}, \ldots, b_{t_s}$ denote the bandwidths, the support set of the kernel function $K$ is $[-1, 1]$, then the estimate of $F$ is

$$\hat{F}(t_1, \ldots, t_s) = \frac{\sum_{i=1}^n Y_i \prod_{j=1}^s K(\frac{t_j - v_{ji}}{b_{t_j}})}{\sum_{i=1}^n \prod_{j=1}^s K(\frac{t_j - v_{ji}}{b_{t_j}})},$$

where $(v_{1i}, \ldots, v_{si})$ denotes points within the range of distance $h$ to point $(t_1, \ldots, t_s)$ [9]. Then, we have the kernel estimates for $f_m, f_l, f_r$,

$$\begin{cases} \hat{f}_m(\underline{X}) = \dfrac{\sum_{i=1}^n Y_i^m \prod_{j=1}^p K(\frac{X_{ji}^m - X_j^m}{b_j^m})K(\frac{X_{ji}^l - X_j^l}{b_j^l})K(\frac{X_{ji}^r - X_j^r}{b_j^r})}{\sum_{i=1}^n \prod_{j=1}^p K(\frac{X_{ji}^m - X_j^m}{b_j^m})K(\frac{X_{ji}^l - X_j^l}{b_j^l})K(\frac{X_{ji}^r - X_j^r}{b_j^r})}, \\ \hat{f}_l(\underline{X}) = \dfrac{\sum_{i=1}^n g(Y_i^l) \prod_{j=1}^p K(\frac{X_{ji}^m - X_j^m}{b_j^m})K(\frac{X_{ji}^l - X_j^l}{b_j^l})K(\frac{X_{ji}^r - X_j^r}{b_j^r})}{\sum_{i=1}^n \prod_{j=1}^p K(\frac{X_{ji}^m - X_j^m}{b_j^m})K(\frac{X_{ji}^l - X_j^l}{b_j^l})K(\frac{X_{ji}^r - X_j^r}{b_j^r})}, \\ \hat{f}_r(\underline{X}) = \dfrac{\sum_{i=1}^n h(Y_i^r) \prod_{j=1}^p K(\frac{X_{ji}^m - X_j^m}{b_j^m})K(\frac{X_{ji}^l - X_j^l}{b_j^l})K(\frac{X_{ji}^r - X_j^r}{b_j^r})}{\sum_{i=1}^n \prod_{j=1}^p K(\frac{X_{ji}^m - X_j^m}{b_j^m})K(\frac{X_{ji}^l - X_j^l}{b_j^l})K(\frac{X_{ji}^r - X_j^r}{b_j^r})}, \end{cases} \tag{3}$$

In this case, we have employed the same vector $(b_1^m, b_1^l, b_1^r, \ldots, b_p^m, b_p^l, b_p^r)$ for the three regression models because our aim is not to estimate such parameters. Nonetheless, in general, different smoothing parameters vectors can also be considered. Also the selection for the kernel function is not emphasized as the estimates are similar numerically for different kernel functions [9].

*Remark 1* In general, an *LR*-fuzzy number is determined by five elements: the shape functions $L$ and $R$, and the center value $m$, two spreads $l$, $r$. In the regression analysis for *LR*-fuzzy data, it may be positive to consider completely the relationship between the five elements of the response *LR*-fuzzy data and the five elements of the explanatory *LR*-fuzzy data.

## 4 A Linearity Bootstrap Test

For the both models in (1) and (2), the residual sum of squares can be defined as

$$SSE = \sum_{i=1}^{n} \delta_2^2(\tilde{Y}_i^T, \widehat{\tilde{Y}_i^T}), \tag{4}$$

where $\tilde{Y}_i^T = (Y_i^m, g(Y_i^l), h(Y_i^r))$, $\widehat{\tilde{Y}_i^T} = (\hat{Y}_i^m, \widehat{g(Y_i^l)}, \widehat{h(Y_i^r)})$, $i = 1, \cdots, n$.

The null hypotheses as follows need to be tested,

$$H_0 : \begin{cases} f_m(\underline{X}) = \underline{Xa}_m^{'} + b_m, \\ f_l(\underline{X}) = \underline{Xa}_l^{'} + b_1, \\ f_r(\underline{X}) = \underline{Xa}_r^{'} + b_r, \end{cases} \tag{5}$$

against the alternative

$$H_1 : f_m(\underline{X}), f_l(\underline{X}), f_r(\underline{X}) \text{ are smooth and non-linear functions.}$$

We use the test statistic

$$T_n = \frac{SSE_0 - SSE_1}{SSE_1}, \tag{6}$$

where $SSE_0$ is the residual sum of squares under $H_0$ according to the model in (1) and $SSE_1$ is the residual sum of squares according to the model in (2), where $\widehat{\tilde{Y}_i^T} = (\hat{Y}_i^m, \widehat{g(Y_i^l)}, \widehat{h(Y_i^r)}) = (\widehat{f_m}(\underline{X}), \widehat{f_l}(\underline{X}), \widehat{f_r}(\underline{X}))$ are the values estimated by means of kernel functions in (3).

*Remark 2* In this paper we use a Gaussian kernel $K(\frac{X-w}{b}) = \frac{1}{\sqrt{2\pi}b} exp(\frac{(X-w)^2}{2b^2})$.

The smoothing parameter $b$ here can be allowed to be fixed since the level of the test is unaffected by this value. Suitable values of $b$ are from 1/n to 1/2 times the range of the $X$-values. Note that the power of the test could be affected by the selection of smoothing parameter [9].

Based on the approaches [2, 3], we generate $B$ bootstrap samples from a bootstrap population fulfilling the null hypothesis in (5) by means of residual approach. The bootstrap statistic is given by

$$T_n^* = \frac{SSE_0^* - SSE_1^*}{SSE_1^*}.$$

The bootstrap algorithm is summarized as follows:
Step 1: Compute the values $\underline{\hat{a}}_m^{'}$, $\underline{\hat{a}}_l^{'}$, $\underline{\hat{a}}_r^{'}$, $\hat{b}_m$, $\hat{b}_l$, $\hat{b}_r$ and $T_n$.
Step 2: Compute the residuals $e_i^m = Y_i^m - \hat{b}_m - \underline{X}_i\underline{\hat{a}}_m^{'}$, $e_i^l = g(Y_i^l) - \hat{b}_l - \underline{X}_i\underline{\hat{a}}_l^{'}$, $e_i^r = h(Y_i^r) - \hat{b}_r - \underline{X}_i\underline{\hat{a}}_r^{'}$.
Step 3: Generate a bootstrap sample of the form
$\{(\underline{X}_1, Z_1^m = \hat{Y}_1^m + e_{i_1}^m, Z_1^l = \widehat{g(Y_1^l)} + e_{i_1}^l, Z_1^r = \widehat{h(Y_1^r)} + e_{i_1}^r), \cdots, (\underline{X}_n, Z_n^m = \hat{Y}_n^m + e_{i_n}^m, Z_n^l = \widehat{g(Y_n^l)} + e_{i_n}^l, Z_n^r = \widehat{h(Y_n^r)} + e_{i_n}^r)\}$,
where $\{i_1, i_2, \ldots, i_n\}$ is random sample of the integers 1 through $n$, $\hat{Y}_i^m = \hat{b}_m + \underline{X}_i\underline{\hat{a}}_m^{'}$, $\widehat{g(Y_i^l)} = \hat{b}_l + \underline{X}_i\underline{\hat{a}}_l^{'}$, $\widehat{h(Y_i^r)} = \hat{b}_r + \underline{X}_i\underline{\hat{a}}_r^{'}$, $i = 1, 2, \cdots, n$ and compute the value of the bootstrap statistic $T_n^*$.
Step 4: Repeat the step 3 a large number $B$ of times to get a set of $B$ estimators, denoted by $\{T_{n1}^*, \cdots, T_{nB}^*\}$.
Step 5: Approximate the bootstrap $p$-value as the proportion of values in $\{T_{n1}^*, \ldots, T_{nB}^*\}$ being greater than $T_n$.

## 5 A Simple Simulation Example

Assume that given a data collection, in which the pulse frequency $Y$, diastolic pressure $X_1$ and systolic pressure $X_2$ for 11 patients with heart disease are recorded. Based on the experts experience we summarize them with the symmetric fuzzy data approach, the artificially processed data given in Table 1. Based on the data in Table 1, we obtain the parameters estimators as follows for a simple case of (1) with $g = h = ln$, where the centers of $Y$ are explained only in terms of centers of $X_1, X_2$, the left spreads of $Y$ are explained only by the left spreads of $X_1, X_2$, and the right spreads of $Y$ are explained only by the right spreads of $X_1, X_2$.
$\hat{a}_m = (0.3337, 0.1648)$, $\hat{b}_m = 21.1603$; $\hat{a}_l = (0.0154, 0.0372)$, $\hat{b}_l = 2.1932$; $\hat{a}_r = (0.0154, 0.0372)$, $\hat{b}_r = 2.1932$ and based on which we compute the residuals $\tilde{\varepsilon} = (\varepsilon^m, \varepsilon^l, \varepsilon^r)$ as the results shown in Table 2. Furthermore, we have considered two gaussian kernels with the smoothing parameters $b_1$ = range$(X_1^m)/2$, $b_2$ = range$(X_2^m)/2$, we obtain

$$SSE_0 = 725.8062, SSE_1 = 1622.0, T_n = -0.5525.$$

Taking a bootstrap sample from the obtained residuals set as a new random sample, based on which $SSE_0^*$ and $SSE_1^*$ can be carried out, and then the corresponding

**Table 1** A set of symmetric triangular fuzzy data for heart disease

| i | Pulse frequency | Diastolic pressure | Systolic pressure |
|---|---|---|---|
| 1 | (56, 12, 12) | (95, 5, 5) | (60, 10, 10) |
| 2 | (66, 6, 6) | (110, 20, 20) | (80, 10, 10) |
| 3 | (73, 17, 17) | (160, 20, 20) | (95, 5, 5) |
| 4 | (91, 21, 21) | (126, 16, 16) | (94, 14, 14) |
| 5 | (63, 9, 9) | (95, 5, 5) | (60, 10, 10) |
| 6 | (85, 15, 15) | (145, 15, 15) | (95, 15, 15) |
| 7 | (69, 6, 6) | (80, 20, 20) | (145, 5, 5) |
| 8 | (86, 14, 14) | (145, 15, 15) | (83, 7, 7) |
| 9 | (87,11, 11) | (150, 40, 40) | (90, 20, 20) |
| 10 | (91, 5, 5) | (159, 21, 21) | (100, 10, 10) |
| 11 | (93, 7, 7) | (130, 20, 20) | (89, 11, 11) |

**Table 2** Computational results of the residuals

| i | $(\varepsilon^m, \varepsilon^l, \varepsilon^r)$ | i | $(\varepsilon^m, \varepsilon^l, \varepsilon^r)$ |
|---|---|---|---|
| 1 | (−6.7498, −0.0033, −0.0033) | 7 | (−2.7523, −0.2794, −0.2794) |
| 2 | (−5.0513, −0.4654, −0.4654) | 8 | (2.7748, 0.4165, 0.4165) |
| 3 | (−17.2083, 0.762, 0.762) | 9 | (0.9527, 0.0767, 0.0767) |
| 4 | (12.3023, 0.5769, 0.5769) | 10 | (0.3014, −0.6324, −0.6324) |
| 5 | (0.2502, −0.291, −0.291) | 11 | (13.7915, −0.3485, −0.3485) |
| 6 | (−0.2028, 0.1879, 0.1897) | | |

$T_n^*$. Repeat above procedure $B = 10000$ times, then we will obtain a set consists of 10000 elements $T_n^*$. The bootstrap $p$-value can be carried out, $p = 0.1346$, thus, for the ordinary nominal significance levels $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$, we have to make rejection of the null hypothesis, there are no obvious linear relationship between response center and explanatory centers, and so is for response spread and explanatory spreads. In general, such results could be affected by the factors such as selected distances, kernel functions as well as the smoothing parameters.

**Conclusion**: In this paper, we consider a linearity test for a special simple case for the model presented in [8]. In which some multiple linear regression model with *LR*-fuzzy response and *LR*-fuzzy explanatory variables proposed by [5] is employed. A multiple nonparametric regression model is estimated, which is employed as a reference in the testing approach. Following [3], a bootstrap procedure is suggested and with which the linearity of the regression model is tested. In the future, it is important to propose a suitable linearity test for the two-variate linear regression model with $n$-dimensional fuzzy random sets in [8].

**Acknowledgments** Financial support from the national natural science foundation of China (NNSFC) with grant number 11271096, is kindly acknowledged.

## References

1. Dimond P, Körner R (1997) Extended fuzzy linear models and least squares estimates. Comput Math Appl 33:15–32
2. Efron, Tibshirani (1993) An Introduction to the Bootstrap. Chapman & Hall
3. Ferraro MB, Colubi A, Giordani P (2010) A linearity test for a simple regression model with LR fuzzy response. In: Combining soft computing and statistical methods. AISC, vol 77, pp 263–271
4. Ferraro MR, Coppi R, Gonzalez-Rodriguez G, Colubi A (2010) A linear regression model for imprecise response. Int J Approximate Reasoning 51:759–770
5. Ferraro MR, Girodani P (2012) A multiple linear regression model for imprecise information. Metrika 75:1049–1068
6. Gil MA (2007) Gonzlez-Rodrguez G, Colubi A, Montenegro M, Testing linear independence in linear models with interval-valued data. Comput Stat Data Anal 51:3002–3015
7. Näther W (2006) Regression with fuzzy random data. Comput Stat Data Anal 51:235–252
8. Wang D, Shi M (2015) Estimation of a simple multivariate linear model with fuzzy random sets. In: Advances in intelligent systems and computing, vol 315, Springer, pp 201–208
9. Wasserman L (2010) All of nonparametric statistics, Springer

# Treat a Weak Dependence Among Causes and Lives in Insurance by Means of Fuzzy Sets

**Dabuxilatu Wang and Tengteng Wang**

**Abstract** In this paper, we apply the copulas and fuzzy sets to approximate the dependencies in causes and lives, where under each cause of decrement the decrement times of the lives are assumed to be weak dependence. We propose utilizing a mixture of both randomness and fuzziness to describe the concept of weak dependence. An application is considered for a general symmetric status of multiple life under dependent causes of decrement.

**Keywords** Weak dependency · Survival function · Copulas

## 1 Introduction

In a multiple life model, the survival times of the lives may be influenced by the common economic/physical environment (such as earthquake, diseases, etc.), which makes the decrement times of lives no longer independent. For example, clinical studies demonstrate that there are some sense of weak positive dependence relations between the decrement times of a married couple [1]. On the other hand, in the multiple decrement model, the decrement causes may be mutually dependent. For instance, the dependency between "death" and "injury". Therefore, there are maybe various unknown dependent relationships among the causes and lives in the concerned insurance models. Carriere [2] proposed dependent decrement theory based on dependent competing risks analysis, in which the survival of the single life under multiple dependent decrement causes is modelled by means of the copulas. A useful

D. Wang (✉)
Department of Statistics, School of Economics and Statistics,
Guangzhou University, No. 230 Waihuanxi Road, Higher Education Mega Center,
Guangzhou 510006, People's Republic of China
e-mail: wangdabu@gzhu.edu.cn

T. Wang
Department of Epidemiology, Gilling School of Public Health,
The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: tengteng@live.unc.edu

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_62

nonparametric estimation on the net survival functions based on the observations of crude survivals was proposed. Kaishev et al. [3] modeled the joint distribution of competing risks survival times using copulas, which improved the dependent decrement theory [2] in its methodology. Dimitrova et al. [4] further consider the same problems in [3] under a causes elimination condition. A popular and unrealistic assumption in a multiple life model is that the considered decrement times of the lives are independent. However, in real world the decrement times of the lives under one of the decrement cause could be dependent in much complicated way, for instance, the way of weak dependence, etc. Ribas [1] pointed out that in most real situation the decrement times of the lives are nearer to the independency. However, here the "nearer to the independency" does not mean a complete independency, there is some sense of weak dependence among the lives from the view of sampling observation.

This paper is organized as follows. In Section one, we introduce some background information. In Section two, some preliminaries around the basic concepts are introduced. In Section three, a simple model for multiple life under dependent decrements is considered. In Section four, we consider an application of the models to multiple life insurance policy.

## 2 Preliminaries

In this section, we will introduce some notions such as weak dependence, survival function and copula.

The dependent multiple decrements models for the case of one life have been investigated extensively ([2, 3]). Here it is assumed that the life (or individual) aged $x \geqslant 0$ may withdraw by any one of the $m$ causes, which means that at birth, the individual is assigned times $T_1, \ldots, T_m$, $0 \leqslant T_j < \infty$, $j = 1, \ldots, m$, representing individual's potential decrement time. We can only observe the $\min(T_1, \cdots, T_m)$, and $T_1, \cdots, T_m$ are unobservable. Real-life actuarial applications indicate that the causes of decrement tend to be dependent. Their *joint distribution function* is denoted by $F(t_1, \cdots, t_m) = P(T_1 \le t_1, \cdots, T_m \le t_m)$, which is assumed to be absolutely continuous. Their *joint survival function* is denoted by $S(t_1, \cdots, t_m) = P(T_1 > t_1, \cdots, T_m > t_m)$, which is absolutely continuous, where $t_j \ge 0$, for $j = 1, \cdots, m$. The *overall survival function* of an individual aged $x \ge 0$ is defined through random variable $\min(T_1, \cdots, T_m)$ as $\mathbb{S}(t) := S(t, \cdots, t) = P(T_1 > t, \cdots, T_m > t) = P(\min(T_1, \cdots, T_m) > t)$. The *crude survival function* $S^{(j)}(t)$ is defined as $S^{(j)}(t) = P(\min(T_1, \ldots, T_m) > t, \min(T_1, \cdots, T_m) = T_j)$, which represents the probability that the observable decrement time exceeds $t$ under the known cause $T_j$, and in practice which can be captured by the sampling observation. The *net survival function* $S^{'(j)}(t)$ is defined as $S^{'(j)}(t) = P(T_j > t)$, and in reality we can not obtain it through a direct observation, only can we observe $\min(T_1, \cdots, T_m)$.

The joint survival function $S(x_1, \cdots, x_n)$ via an appropriate copula [5, 6] $\overline{C}$ is said to be the survival copula of the random vector $(X_1, \cdots, X_n)$.

**Lemma 2.1** ([2, 3]) *The joint survival function can be expressed by the unique survival copula, i.e.,* $S(x_1, \cdots, x_n) = \overline{C}(S^{'(1)}(x_1), \cdots, S^{'(n)}(x_n))$.

**Lemma 2.2** ([2]) *The crude survival function possesses following relation with the joint survival function,* $S^{(j)}(t) = \int_t^{+\infty} -S_j(r, \cdots, r)dr$, *where* $S_j(r, \cdots, r) = \frac{\partial}{\partial t_j} S(t_1, \cdots, t_n)\Big|_{t_k = r, \forall k}$.

The Kendall's $\tau$ is defined as the probability of concordance minus the probability of discordance: $\tau := P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$, where $(X_1, Y_1)$ and $(X_2, Y_2)$ are independent and identically distributed random vectors. Furthermore, If $X$ and $Y$ are continuous random variables whose copula is C, then the population version of Kendall's $\tau$ for $X$ and $Y$ can be given by $\tau = 4 \int\int_{[0,1]\times[0,1]} C(u, v)dC(u, v) - 1 = 3 - 4 \int_0^1 K(t)dt$, where $K(t)$ is the distribution function of the random variable $C(u, v)$, and the random variables $u, v$ are uniformly distributed on [0, 1]. Kendall's $\tau$ can be used for description of the strength of dependence in copulas. For some special copulas such as Gumbel's copula, Clayton's copula and Frank's copula, the independence between random variables concerned with the copulas can be equivalent to that their corresponding Kendall's $\tau$ taken value of zero [3, 5–7].

The notion of the weak dependence aforementioned for describing the lower strength of dependence of the decrement times of lives is hard to define in precise with the probability setting, since here the word "weak" is a vague (fuzzy) description. As that proposed in [7] this concept could be expressed by some special fuzzy set $\tilde{k}$ (a membership function) defined on a domain $[-\varepsilon, \varepsilon]$, i.e. $\tilde{k} : [-\varepsilon, \varepsilon] \to [0, 1]; \tau \to \tilde{k}(\tau) \in [0, 1]$, the domain is assumed to be the set of possible values (close to zero) of the Kendall's $\tau$ that defined through the three kinds of copulas aforementioned, where $\varepsilon$ is a small positive real number. In other words, the notion "weak dependence" means a fuzzy Kendall's $\tau$ symmetrically around zero, for which the membership function is not unique, one can construct the membership functions depend on the real context of the dependence. For instance, an expert may think that based on his/her own experiences the considered two random variables are weak dependent with different levels if their Kendall's $\tau$ takes values in the range $[-0.05, 0.05]$, and the fuzzy Kendall's $\tau$ symmetrically around zero is selected to be a fuzzy set $\tilde{k}$, a trapezoidal fuzzy set on the interval $[-0.05, 0.05]$, or simply some discrete fuzzy set as $\tilde{k} = \{(\tau, \mu) : \tau \in [-\varepsilon, \varepsilon], \mu \in [0, 1]\}$, which represent kinds of expressions for the weak dependence.

## 3 A Simple Model for Multiple Life Under Dependent Decrements

Consider $n$ lives aged $x_1, x_2, \cdots, x_n$, where each of the $n$ lives is exposed to $m$ dependent causes of decrement denoted by $I = \{1, 2, \cdots, m\}$. Then, all possible decrement times are the $mn$ different decrement times shown in Table 1. The assumption

**Table 1** *mn* decrement variables

| | $x_1$ | $x_2$ | $\dots$ | $x_n$ |
|---|---|---|---|---|
| $T_1$ | $T_1(x_1)$ | $T_1(x_2)$ | $\dots$ | $T_1(x_n)$ |
| $T_2$ | $T_2(x_1)$ | $T_2(x_2)$ | $\dots$ | $T_2(x_n)$ |
| $T_3$ | $T_3(x_1)$ | $T_3(x_2)$ | $\dots$ | $T_3(x_n)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\dots$ | $\vdots$ |
| $T_m$ | $T_m(x_1)$ | $T_m(x_2)$ | $\dots$ | $T_m(x_n)$ |

of dependencies of the $m$ causes makes the $m$ decrement time variables in each column of Table 1 become a random vector of $m$-dimension with dependent components, $\overline{T}_{x_j} := (T_1(x_j), \cdots, T_m(x_j)), j = 1, \cdots, n$. and we are able to observe only $\min \overline{T}_{x_j} := \min\{T_1(x_j), \cdots, T_m(x_j)\}, j = 1, \cdots, n$. Based on the description about the relation of individuals in previous section, we may assume that under each cause the decrement times of the $n$ lives are also dependent, then the $n$ decrement time variables of each row in Table 1 can be viewed as a latent random vector of $n$-dimension with dependent components, $\overline{T}_i := (T_i(x_1), \cdots, T_i(x_n)), i = 1, \cdots, m$.

Based on Lemmas 2.1 and 2.2, it holds

$$\frac{d}{dt}S_l^{(j)}(t) = \overline{C}_{jl}(S_1^{'(1)}(t), \cdots, S_n^{'(1)}(t), \cdots, S_1^{'(m)}(t), \cdots, S_n^{'(m)}(t)) \times \frac{dS_l^{'(j)}(t)}{dt}.$$

where $\overline{C}_{jl}(u_{11}, \cdots, u_{mn}) = \frac{\partial}{\partial u_{jl}}\overline{C}(u_{11}, \cdots, u_{mn})$ for $j = 1, \cdots, m, l = 1, \cdots, n$.

### 3.1 The Case Where the Decrement Times of the Lives Are Weak Dependent

Based on the notion of the weak dependence of two random variables introduced in Sect. 2, we consider two lives $x$, $y$ under dependent decrement causes $\{1, 2, \cdots, m\}$. Assume that for each cause $i \in \{1, 2, \cdots, m\}$, the latent continuous decrement times variables $T_i(x)$, $T_i(y)$ are weak dependent, then their relation could be described by some fuzzy Kendall's tau $\tilde{k}_i$ defined on the interval $[-\varepsilon_i, \varepsilon_i]$, the set of possible values of the crisp Kendall's $\tau_i$ symmetrically around zero concerning the three kinds of copulas aforementioned, i.e.

$$\tilde{k}_i = \{(\tau_i, \mu_i) : \tau_i \in [-\varepsilon_i, \varepsilon_i], \mu_i \in [0, 1]\},$$

note that there exists some positive invertible function $h_i$ such that $\tau_i = h_i(\theta_i)$, where $\theta_i$ is the parameter of the copula $C_i(u, v; \theta_i)$ that is used for modeling the dependence between $T_i(x)$ and $T_i(y)$ and determine the Kendall's $\tau_i$, and $\theta_i = h_i^{-1}(\tau_i)$, $i \in \{1, 2, \cdots, m\}$. We propose to define the corresponding fuzzy parameter $\tilde{\theta}_i$ of fuzzy Kendall's tau $\tilde{k}_i$ symmetrically around zero as a set

$$\tilde{\theta}_i := \{(h_i^{-1}(\tau_i), \mu_i) : h_i^{-1}(\tau_i) = \theta_i \in (0, h_i^{-1}(\varepsilon_i)), \mu_i \in [0, 1]\},$$

and the copula with fuzzy parameter $\tilde{\theta}_i$ as a set

$$C_i(u, v; \tilde{\theta}_i) := \{C_i(u, v; \theta_i) : (\theta_i, \mu_i) \in \tilde{\theta}_i\}.$$

We also note that we could only observe $\min \overline{T}_x$, $\min \overline{T}_y$, the obstacle of identifiability make us not able to estimate the parameters $\theta_i$ intuitively based on observations [3], sometimes we have to assign special values to them based on predefined values of $\tau_i$.

Assume that there exists a copula $C$ which could be used for modeling the dependence among the causes of decrement, and moreover, assume that the compositions of copulas $C(C_1, \cdots, C_m)$ could be used to model the dependencies existed in both lives and causes.

**Theorem 3.1** *If the latent continuous decrement times $T_i(x)$, $T_i(y)$ are weak dependent for each i, $i = 1, 2, \cdots, m$, then the observable decrement times* $\min \overline{T}_x$, $\min \overline{T}_y$ *are also weak dependent.*

*Proof* We denote the joint distribution and marginal distribution functions of ($\min \overline{T}_x$, $\min \overline{T}_y$) by $F_{\min \overline{T}_x, \min \overline{T}_y}$, $F_{\min \overline{T}_x}$ and $F_{\min \overline{T}_y}$, respectively. By Sklar theorem [5, 6] there exists a copula $C$ such that $F_{\min \overline{T}_x, \min \overline{T}_y}(t_1, t_2) = C(F_{\min \overline{T}_x}(t_1), F_{\min \overline{T}_y}(t_2)) = C(1 - S_x(t_1, \cdots, t_1), 1 - S_y(t_2, \cdots, t_2)) = C(1 - \overline{C}_1(S_{T_1(x)}(t_1), \cdots, S_{T_m(x)}(t_1)), 1 - \overline{C}_2(S_{T_1(y)}(t_2), \cdots, S_{T_m(y)}(t_2)))$, where $S_x$, $S_y$ are survival functions and $\overline{C}_1$, $\overline{C}_2$ are survival copulas. When
$\overline{C}_1(S_{T_1(x)}(t_1), \cdots, S_{T_m(x)}(t_1)) = S_{T_1(x)}(t_1) = \overline{C}_1(S_{T_1(x)}(t_1), 1, \cdots, 1)$,
$\overline{C}_2(S_{T_1(y)}(t_2), \cdots, S_{T_m(y)}(t_2)) = S_{T_1(y)}(t_2) = \overline{C}_2(S_{T_1(y)}(t_2), 1, \cdots, 1)$, we have $F_{\min \overline{T}_x, \min \overline{T}_y}(t_1, t_2) = C(1 - S_{T_1(x)}(t_1), 1 - S_{T_1(y)}(t_2))$, which means that the copula $C$ is a copula modeling a weak dependent relation. ■

If the conditions of Theorem 3.1 are fulfilled, then using the copulas' compositions we are able to approximate the weak dependence between $\min \overline{T}_x$ and $\min \overline{T}_y$ extensively, i.e. constructing a set of copulas

$$\{C(F_{\min \overline{T}_x}, F_{\min \overline{T}_y}; \theta_{min}), (\theta_{min}, \mu_{min}) \in \tilde{\theta}_{min}\},$$

where $F_{\min \overline{T}_x}$, $F_{\min \overline{T}_y}$ are the distribution, expressed by some copulas, of $\min \overline{T}_x$, $\min \overline{T}_y$, respectively. $\tilde{\theta}_{min}$ is the corresponding fuzzy parameter induced from the fuzzy Kendall's tau symmetrically around zero for describing the weak dependency between $\min \overline{T}_x$ and $\min \overline{T}_y$.

## 4 An Application to Multiple Life Insurance Policy

As illustrated in [8], we define the status $u$ by means of the residual non-decrement time $T_u$. By ${}_tp_u$ we denotes the conditional probability that the status $u$ still exists at time $t$, given that the status $u$ existed at time 0. A general symmetric status denoted by $u(k)$ means that the status $u(k)$ fails at the $(n-k+1)$th withdrawal. We now consider the computation of the non-decrement probability ${}_tp_{u(k)}$ of the general symmetric status for the observable decrement time variables $\min \overline{T}_{x_1}$, $\min \overline{T}_{x_2}$, $\cdots$, $\min \overline{T}_{x_n}$ of the $n$ lives, which depends upon the joint distribution of the random vector $(\min \overline{T}_{x_1}, \min \overline{T}_{x_2}, \cdots, \min \overline{T}_{x_n})$. Note that, the non-decrement probability of $\min \overline{T}_{x_j}$ can be derived by some (survival) copula $C_j$ that is used for modeling the dependencies of the decrement causes $j, j = 1, 2, \cdots, n$. i.e.

$$\begin{aligned} {}_tp_{\min \overline{T}_{x_j}} &:= S_{\min \overline{T}_{x_j}}(t) = P(\min\{T_1(x_j), T_2(x_j), \cdots, T_m(x_j)\} > t) \\ &= C_j(S_{T_1(x_j)}(t), \cdots, S_{T_m(x_j)}(t)). \end{aligned}$$

where $S_{T_i(x_j)}(t) := S_{x_j}^{'(i)}(t)$ is the net survival function of individual $x_j$ under decrement cause $i$, which need to be estimated via solving differential equations aforementioned based on data of the crude survival functions $S_{x_j}^{(i)}(t)$.

Let the event $\{\min \overline{T}_{x_j} > t\}$ be denoted by ${}_tB_j$, $j = 1, 2, \cdots, n$. Set${}_tS_k := \sum_{J_k \in CL(n,k)} P({}_tB_{j_1}, {}_tB_{j_2}, \cdots, {}_tB_{j_k})$, $k = 1, 2, \cdots, n$. Here $J_k := \{j_1, j_2, \cdots, j_k\}$ and $CL(n, k)$ denotes the class of all $C_n^k$ subsets of $\{1, 2, \cdots, n\}$ with $k$ different elements, $k = 1, 2, \cdots, n$. By the Sklar theorem $P({}_tB_{j_1}, {}_tB_{j_2}, \cdots, {}_tB_{j_k})$ can be derived by some copula $C$. Then, ${}_tS_k := \sum_{J_k \in CL(n,k)} C(S_{\min \overline{T}_{x_{j_1}}}(t), \cdots, S_{\min \overline{T}_{x_{j_k}}}(t))$, $k = 1, 2, \cdots, n$, and the Schuette-Nesbitt formula [8], ${}_tp_{u(k)} = \sum_{w=k}^{n} (-1)^{w-k} C_{w-1}^{k-1} \sum_{J_w \in CL(n,w)} C(S_{\min \overline{T}_{x_{j_1}}}(t), \cdots, S_{\min \overline{T}_{x_{j_w}}}(t))$.

In the following we consider a 5-lives group under 3 dependent causes of decrement. The copulas for modeling the dependency among lives could be

(1) Frank copula (Fc): $C(u_1, \cdots, u_n) = -\frac{1}{\theta} ln[1 + \frac{\prod_{i=1}^{n}(e^{-\theta u_i} - 1)}{(e^{-\theta} - 1)^{n-1}}]$;

(2) Generalized Clayton copula (GCc): $C(u_1, \cdots, u_n) = [u_1^{-\theta} + \cdots + u_n^{-\theta} - n + 1]^{-\frac{1}{\theta}}$;

(3) Independent copula (Ic): $C(u_1, \cdots, u_n) = \prod_{i=1}^{n} u_i$. Where $\theta$ is the copula's parameter taken values in $(0, +\infty)$.

Assume that a group of five lives aged $x_1, \cdots, x_5$ respectively are exposed to the dependent decrement causes $d, w, i$, the probabilities of decrement can refer to [8, 9], where $d, w, i$ represent "death", "withdraw during occupation", "injury", respectively.

Assume that the dependency of the causes $d, w, i$ can be modelled by copula GCc or Fc, then solving the non-linear differential equations that model the relation between the crude and net survival functions via the *Mathematica* built-in functions `NDSolve`, we may illustrate solution net survival with the curves shown in Fig. 1.

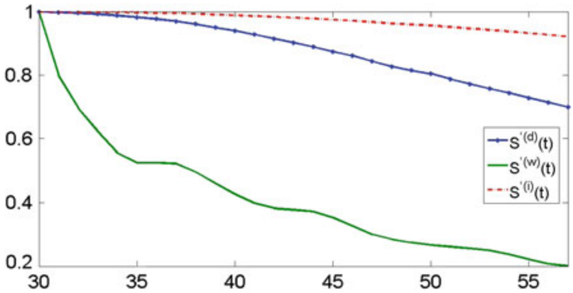


**Fig. 1** The curves of the net survival under Clayton copula

Since each individual of the 5-lives could be affected by all causes $d, w, i$, the copulas are assumed to be same, $C_1 = C_2 = C_3 = C_4 = C_5$, for simplicity, here we choose the common copula to be GCc and assign its parameter $\theta = 2$, as the strength of dependence among the three causes seems stronger than that of independence relation. Then we obtain the probabilities $p_{\min \overline{T}_{x_1}} = 0.9925, p_{\min \overline{T}_{x_2}} = 0.9573$, $p_{\min \overline{T}_{x_3}} = 0.9228$, $p_{\min \overline{T}_{x_4}} = 0.9187$, $p_{\min \overline{T}_{x_5}} = 0.9100$, where $x_1 = 35$, $x_2 = 38$, $x_3 = 40$, $x_4 = 42$, $x_5 = 45$.

The probabilities $p_{u(k)}$, $k = 1, 2, 3, 4, 5$ that each general symmetric status exceeds 1 year under the copulas aforementioned can be carried out. The dependence among lives is assumed to be weak dependent, where the parameter $\theta$ of the concerned copulas to be valued in [0.01, 1] with membership degree $\mu(\theta) = 1 - \theta$. Results are shown in Table 2.

The results from Table 2 show that for the Fc, GCc, each $p_{u(i)}$, $i = 1, 2, 3, 4$ is a decreasing function of $\theta$, and $p_{u(5)}$ is a increasing function of $\theta$, $\theta \in [0.01, 1]$. The interval value of each $p_{u(i)}$, $i = 1, 2, 3, 4$ for the Fc and GCc includes all of values with all different membership values, it accurately describe the characteristics of the weak dependence by means of fuzzy sets. The right end point of each interval with membership 0.99 seems to almost equal to the values of the Ic, which means that using Ic maybe appropriate, meanwhile the other values in each interval with some membership values seems to smaller than the values of the Ic, which means that using Ic may bring an underestimation on risks. One may accurately estimate the risk depending on some confident membership value. $u(5)$ is a join survival status for 5 lives, values in the interval of $p_{u(5)}$ for Fc, GCc, are bigger than the values of the Ic, it may overestimate the risks with some sense of risk defence.

**Table 2** Comparisons of probabilities $p_{u(k)}$, $k = 1, 2, 3, 4, 5$ under copulas ($\theta \in [0.01, 1]$, $\mu(\theta) = 1 - \theta$, the superscripts represent the times number 9 repeated, e.g. $0.9^4 87 = 0.999987$.)

| | $p_{u(1)}$ | $p_{u(2)}$ | $p_{u(3)}$ | $p_{u(4)}$ | $p_{u(5)}$ |
|---|---|---|---|---|---|
| Fc | $[0.9^4 87, 0.9^6 6]$ | $[0.9^3 48, 0.9^4 4]$ | $[0.9^2 368, 0.9^2 765]$ | [0.95085, 0.96138] | [0.70083, 0.71579] |
| GCc | $[0.9^4 79, 0.9^6 6]$ | $[0.9^3 16, 0.9^4 4]$ | $[0.9^2 077, 09^2 762]$ | [0.94114, 0.96122] | [0.70102, 0.72876] |
| Ic | $0.9^6 7$ | $0.9^4 41$ | $0.9^2 7682$ | 0.9614866 | 0.7006897 |

**Conclusions:** We consider the case where the decrement times of the considered multiple life are weak dependent under each decrement cause, while the decrement causes are also dependent. Some multivariate copulas combined with fuzzy Kendall' $\tau$ are employed to model such dependencies.

**Acknowledgments** This research is supported by the NNSF of China under grant number No. 11271096. Their financial aid is greatly appreciated.

## References

1. Ribas C, Marín-Solano J, Alegre A (2003) On the computation of the aggregate claims distribution in the individual life model with bivariate dependencies. Insur Math Econ 32:201–215
2. Carriere JF (1994) Dependent decrement theory. Trans Soc Actuaries 46:45–74
3. Kaishev VK, Dimitrova DS, Haberman S (2007) Modelling the joint distribution of competing risks survival times using copula functions. Insur Math Econ 41:339–361
4. Dimitrova DS, Haberman S, Kaishev VK (2013) Dependent competing risks: cause elimination and its impact on survival. Insur Math Econ 53:464–477
5. Durante F, Sempi C (2015) Principles of copula theory. Chapman and Hall/CRC
6. Nelsen R (1999) An introduction to copulas. Springer, New York
7. Hryniewicz O (2010) On testing fuzzy independence application in quality control. In: Borgelt, C et al. (eds.) Combining soft computing & statistical methods. AISC vol 77. Springer, Berlin, Heidelburg, pp 337–344
8. Gerber HU (1997) Life insurance mathematics. Springer
9. Bowers NL, Gerber HU, Hickman JC, Jones DA, Nesbitt CJ (1997) Actuarial mathematics, 2nd edn. The society of Actuares, Schaumburg, Illinois

# A Portfolio Diversification Strategy via Tail Dependence Clustering

**Hao Wang, Roberta Pappadà, Fabrizio Durante and Enrico Foscolo**

**Abstract** We provide a two-stage portfolio selection procedure in order to increase the diversification benefits in a bear market. By exploiting tail dependence-based risky measures, a cluster analysis is carried out for discerning between assets with the same performance in risky scenarios. Then, the portfolio composition is determined by fixing a number of assets and by selecting only one item from each cluster. Empirical calculations on the EURO STOXX 50 prove that investing on selected assets in trouble periods may improve the performance of risk-averse investors.

## 1 Introduction

In recent years, financial markets have been characterized by an increasing globalization and a complex set of relationships among asset returns. Moreover, it has been recognized that the linkages among different assets vary across time and that their strength tends to increase especially during crisis periods. The presence of a stronger dependence when markets are experiencing losses, is of utmost interest from a risk manager perspective. In fact, it has been recognized that investors can reduce the risk of their portfolios through diversification, i.e. allocating their investments in various classes and/or categories that would move in different ways in response to the same event.

H. Wang
School of Economics, Jilin University, Changchun 130012, China
e-mail: haowang@jlu.edu.cn

R. Pappadà
Department of Economics, Business, Mathematics and Statistics "Bruno De Finetti",
University of Trieste, 34127 Trieste, Italy
e-mail: rpappada@units.it

F. Durante · E. Foscolo (✉)
Faculty of Economics and Management, Free University of Bozen-Bolzano, Bolzano, Italy
e-mail: enrico.foscolo@unibz.it

F. Durante
e-mail: fabrizio.durante@unibz.it

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_63

In order to provide a suitable diversification of a portfolio that takes into account the occurrence of extreme scenarios, various clustering techniques for multivariate time series have been proposed in the literature, mainly based on measures of association like Pearson correlation coefficient (see, e.g., [13]). Recently, such techniques have also been applied in order to group financial time series that are similar in extreme scenarios by using tail dependence coefficients (see, e.g., [2, 3] and [7]), or conditional measures of association, like Spearman's correlation, as done in [6]. For an alternative approach, see also [9, 10].

The aim of this contribution is to exploit recent tail-dependence clustering methods in order to select a weighted portfolio in a group of assets. In particular, it will be shown how the adoption of fuzzy clustering methodology (see, e.g., [8] and references therein) may provide some advantages in terms of both performance and computational tractability of the model.

## 2 The Clustering Procedure

Several clustering procedures are based on the choice of a suitable dissimilarity measure that expresses the relations among the financial time series of the asset returns under consideration. Following previous approaches, we present here a procedure to group time series based on their tail behaviour, as done in [6]. This methodology is summarized below.

Consider a matrix of $d$ financial time series $(x_{it})_{t=1,\dots,T}$ $(i = 1, 2, \dots, d)$ representing the log–returns of different financial assets. We assume that each time series $(x_{it})_{t=1,\dots,T}$ is generated by the stochastic process $(\mathbf{X}_t, \mathcal{F}_t)$ such that, for $i = 1, \dots, d$,

$$X_{it} = \mu_i(\mathbf{Z}_{t-1}) + \sigma_i(\mathbf{Z}_{t-1})\varepsilon_{it}, \tag{1}$$

where $\mathbf{Z}_{t-1}$ depends on $\mathcal{F}_{t-1}$, the available information up to time $t - 1$, and the innovations $\varepsilon_{it}$ are distributed according to a distribution function $F_i$ for each $t$. Moreover, the innovations $\varepsilon_{it}$ are assumed to have a constant conditional distribution $F_i$ (with mean zero and variance one, for identification) such that for every $t$ the joint distribution function of $(\varepsilon_{1t}, \dots, \varepsilon_{dt})$ can be expressed in the form $C(F_1, \dots, F_d)$ for some copula $C$. Such a general model includes many multivariate time series models presented in the literature (see, for instance, [14]).

Then the following steps can be performed in order to divide the time series into sub-groups such that elements in each sub-group have strong tail dependence between each other.

1. Choose a copula-based time series model in order to describe separately the marginal behavior of each time series and the link between them.
2. Estimate a (pairwise) tail dependence measure among all the time series.
3. Define a dissimilarity matrix by using the information contained in the tail dependence matrix.

4. Apply a suitable cluster algorithm for grouping time series according to the tail behavior.

Steps 1–3 described above have been discussed in details in [6]. Here (and in the following illustration), these steps are specified in the following way:

1. We fit an appropriate ARMA-GARCH model to each univariate time series and, using the estimated parameters, we construct the standardized residuals that are helpful in determining the joint distribution of the innovations.
2. As a measure of tail dependence, we use the conditional Spearman's correlation $\rho_\alpha$ that expressed the Spearman's correlation between two random variables $X$ and $Y$ given that they are both under their $\alpha$–quantile (here, $\alpha = 0.10$). The estimation is based on the procedure described in [4, 5].
3. Once the conditional Spearman's correlation has been computed for all pairs extracted from the time series, we transform it through a monotonic function $f$ in such a way that the obtained dissimilarity between two time series is small when their tail dependence is high, and monotonically increases when their tail dependence decreases. Thus, for $i, j = 1, \dots, d$, we define $\Delta = \left(\Delta_{ij}\right)$ whose elements are given by

$$\Delta_{ij} = \sqrt{2(1 - \hat{\rho}_\alpha^{ij})}, \tag{2}$$

   where $\hat{\rho}_\alpha^{ij}$ is the conditional Spearman's correlation between time series $i$ and $j$.

Starting from the dissimilarity matrix defined in (2), we can perform a cluster analysis by different techniques. Here we focus on a fuzzy clustering algorithm, i.e. the *fanny algorithm* by [12], since it allows to quantify the degree of membership of an object to the different clusters by means of a coefficient, which ranges from 0 to 1. In order to determine the optimal number $k$ of clusters, we use the average silhouette index [11], which reflects the within-cluster compactness and between-cluster separation of a clustering.

Fanny algorithm aims to minimize the objective function

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} m(i,v)^r m(j,v)^r \Delta_{ij}}{2 \sum_{j=1}^{n} m(j,v)^r}$$

where $n$ is the number of involved time series, $k$ is the number of clusters, $r > 1$ is the membership exponent (usually, $r = 2$), $m(i, v)$ the membership of time series $i$ to cluster $v$, and $\Delta_{ij}$ is the dissimilarity between the time series $i$ and $j$. The algorithm returns the membership degree of each time series $i$ to any cluster. Obviously, if a crisp assignment of each time series to only one cluster is necessary, then one could proceed according to the highest membership degree.

## 3 The Portfolio Selection

Once the cluster analysis is carried out for identifying assets with the same performance during risky scenarios, a portfolio selection procedure can be implemented by fixing the number of assets per portfolio equal to the number of clusters, and by selecting only one item from each cluster. The rationale is that, since assets in different clusters are weakly associated with each other (in risky periods), then they form a well-diversified portfolio. This idea has been used, for instance, in [2, 3] and is slightly modified here by exploiting the advantages of fuzzy clustering.

Specifically, suppose that $n$ time series have been classified by means of the procedures described in Sect. 2 into $k \geq 2$ groups. Let $m(i, v)$ be the membership degree of time series $i$ to cluster $C_v$. The selection algorithm goes as follows:

**The portfolio selection algorithm**

1. Fix $T \in [0, 1]$, which represents a cut-off value for the degree of membership to a cluster.
2. For $i = 1, 2, \ldots, n$, assign the time series $i$ to the cluster $C_v$ if it holds that $m(i, v) = \max_{v'=1,\ldots k} m(i, v')$.
3. For each cluster $C_v$ $(v = 1, \ldots k)$, remove the element $j$ in $C_v$ provided that $m(j, v) < T$. The resulting clusters are denoted by $D_v$ $(v = 1, \ldots, k)$. Notice that some $D_v$ can be empty.
4. Determine all possible portfolios composed by (at most) $k$ assets obtained by selecting exactly one asset from each element of $\{D_1, \ldots, D_k\}$.
5. For these portfolios, calculate the optimal weights assigned to each of its assets by Minimum Conditional-Value-at-Risk (CVaR) strategy.
6. Select the Minimum CVaR portfolio with the lowest CVaR value.

Some comments are needed here.

Step 3 guarantees that we only focus on those assets that can be assigned to a given cluster with a membership degree larger than $T$. It avoids the selection of assets that are likely to be associated with more than one cluster (and, hence, tend to downgrade the effects of diversification).

Step 4 is usually computationally expensive; however, the computational burden can be limited by a careful selection of the cut-off value $T$. In particular, this aspect highlights the main difference between the proposed algorithm and the methodology discussed in [2].

Step 5 suggests a portfolio selection procedure that focuses on extreme events and, hence, is coherent with the tail dependence approach developed here (see also [3]). Specifically, the procedure optimizes the CVaR, defined as the expected loss exceeding $VaR_\beta$ (for more details, see [15]). Below, we set $\beta = 0.10$.

For the illustration of the algorithm, we consider time series related to EURO STOXX 50 stock index and its components in the period from January 2, 2003 to July 31, 2011. Moreover, as out-of-sample period, we will show the performance of our procedure in the period from August 1, 2011 to September 9, 2011. The period

**Table 1** Cluster composition of the EURO STOXX 50 constituents by using conditional Speaman's correlation $\rho_\alpha$ with $\alpha = 0.1$ and fanny algorithm. The assets whose maximal membership degree is smaller than 0.90 are denoted in bold

| Cluster | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | D.DTEX | E.IND | D.SAPX | F.EI | D.BAYX | F.UBL | D.BMWX |
| | **F.CRFR** | **D.RWEX** | | | | | |
| 2 | F.SQ.F | F.FTEL | F.OR.F | H.ASML | F.EX.F | F.BSN | |
| 3 | E.BBVA | D.ALVX | H.ING | D.MU2X | E.SCH | F.TAL | F.BNP |
| | D.BASX | E.REP | I.ENEL | I.ENI | **D.DBKX** | **F.SGE** | |
| 4 | E.TEF | F.GSZ | H.MT | M.NOK1 | CRGI | D.EONX | F.AIR |
| | B.ABI | **E.IBE** | | | | | |
| 5 | F.QT.F | F.GOB | H.PHIL | D.SIEX | I.ISP | H.UNIL | I.UCG |
| | **I.G** | **F.MIDI** | **D.DAIX** | **F.LVMH** | **F.DG.F** | **D.VO3X** | |

has been selected due to the fact that EURO STOXX 50 was experiencing severe losses (see Fig. 2).

We preliminary apply a univariate Student-$t$ ARMA(1,1)-GARCH(1,1) model to each time series of log–returns of the constituents of the index to remove autocorrelation and heteroscedasticity from the data and compute the standardized residuals.

Then, we compute the conditional Spearman's $\rho_\alpha$ (here we select $\alpha = 0.10$) for all pairs of times series. By means of the procedures illustrated in Sect. 2, we determine a dissimilarity matrix and apply the fanny algorithm. According to the average silhouette index, the optimal number of cluster, $k$, is set equal to 5 (we run different algorithms with $k = 2, 3, \ldots, 8$).

Table 1 presents the cluster composition of the portfolio, when each asset is assigned to a cluster in a crisp way. Moreover, we highlighted in bold all the assets whose maximal membership degree is smaller than $T = 0.90$.

Thus, we run the portfolio selection algorithm by considering all the assets (i.e. by setting $T = 0$) or by considering the assets whose maximal membership degree is larger than $T = 0.90$). All the possible 82134 portfolios composed by 5 assets, such that each asset belongs to a different cluster, are calculated and visualized in Fig. 1, where the 25872 possible portfolios obtained by adopting the threshold $T = 0.90$ are colored in grey. As can be seen, the minimal CVaR portfolios generated by the algorithm with $T = 0$ and $T = 0.90$ coincide; however, the latter is obtained under a smaller computational effort.

In order to verify the performance of the methodology in an out-of-sample comparison, we consider the period from August 1, 2011 to September 9, 2011 as out-of-sample period, and compare the performance of the minimum CVaR portfolios obtained from our algorithm (with $T = 0$ and $T = 0.90$) with, respectively, the minimum variance portfolio and the minimum CVaR portfolio built from the whole set of assets, the equally weighted portfolio (obtained by assigning the same weight to

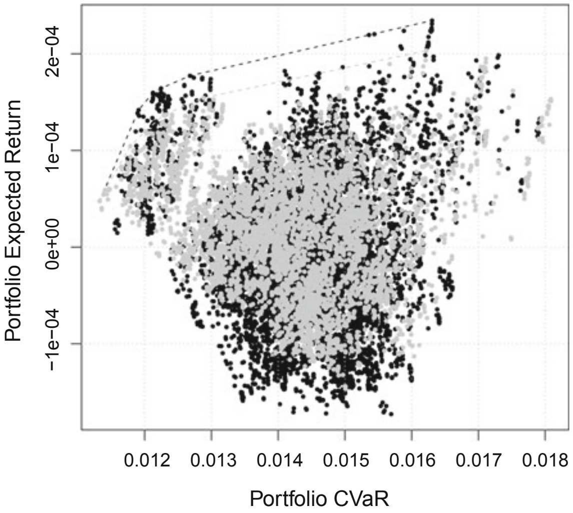


**Fig. 1** Portfolio CVaR–Portfolio Expected Return plot of 5-asset portfolios generated at Step 4 of the portfolio selection algorithm. i highlights the portfolio frontier obtained from our algorithm with $T = 0$ (*black*) and $T = 0.90$ (*gray*)

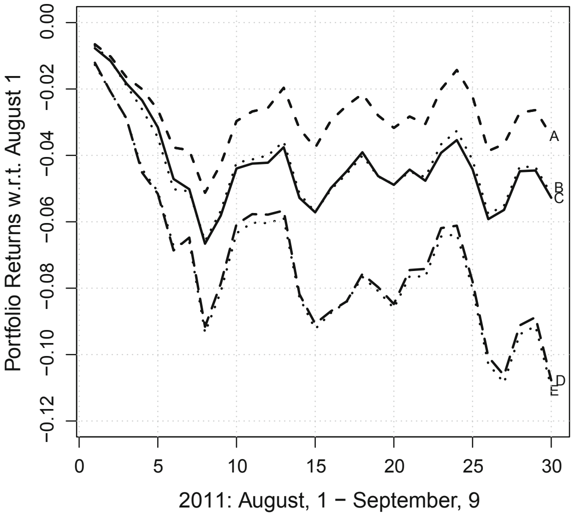


**Fig. 2** Out-of-sample performance of the following portfolios: *A* minimum CVaR portfolio produced by our algorithm, *B* minimum variance portfolio from all 50 assets, *C* minimum CVaR portfolio from all 50 assets, *D* equally weighted portfolio, *E* EURO STOXX 50 index

each asset) and the benchmark index EURO STOXX 50. As it can be seen in Fig. 2, the performance of the portfolios selected from the proposed algorithm is better than the benchmark and outperforms the global minimum variance portfolio. This seems to confirm the idea that, when markets are experiencing a period of losses, a diversification strategy could be beneficial.

## 4 Conclusions

We have introduced a procedure aiming at selecting a portfolio from a group of assets in such a way that the assets are diversified in their tail behavior. The procedure exploits some features of fuzzy clustering algorithms. It is intended to be used by an investor to have more insights into the relationships among different assets in crisis periods.

Although these preliminary findings are promising, further analysis is necessary to assess the validity of the procedures. First, more benchmark datasets should be analyzed to assess the real usefulness of the proposed algorithm. Second, different tail dependence measures and/or clustering procedures (in particular, fuzzy *c*–medoids algorithms [1]) should be considered. Finally, as kindly suggested by one of the reviewers, in order to mitigate the computational burden, it could be also convenient to rank all the possible portfolios according to the sum of the membership degrees of their components and, hence, select the top *p* portfolios (*p* should be decided by the user) for further analysis. All these aspects will be the object of future investigations.

**Acknowledgments** The first author acknowledges the support of the Major Program of the National Social Science Foundation of China (No. 15ZDA017), and the support of Jilin University via the "Fundamental Research Funds for the Central Universities" (No. 450060522110) and via "Young Academic Leaders Training Program" (No. 2015FRLX07). The second author acknowledges the support of the Department of Economics, Business, Mathematics and Statistics "Bruno De Finetti" (University of Trieste, Italy), via the project "FRA". The third and fourth author acknowledge the support of the Faculty of Economics and Management (Free University of Bozen-Bolzano, Italy), via the project "COCCO".

## References

1. Coppi R, D'Urso P, Giordani P (2006) Fuzzy $C$-Medoids clustering models for time-varying data. In: Bouchon-Meunier B, Coletti G, Yager R (eds) Modern information processing: from theory to applications. Elsevier Science, Amsterdam, pp 195–206
2. De Luca G, Zuccolotto P (2011) A tail dependence-based dissimilarity measure for financial time series clustering. Adv Data Anal Classif 5(4):323–340
3. De Luca G, Zuccolotto P (2015) Dynamic tail dependence clustering of financial time series. Stat Pap. doi:10.1007/s00362-015-0718-7
4. Dobrić J, Frahm G, Schmid F (2013) Dependence of stock returns in bull and bear markets. Depend Model 1:94–110
5. Durante F, Jaworski P (2010) Spatial contagion between financial markets: a copula-based approach. Appl Stoch Models Bus Ind 26(5):551–564
6. Durante F, Pappadà R, Torelli N (2014) Clustering of financial time series in risky scenarios. Adv Data Anal Classif 8:359–376
7. Durante F, Pappadà R, Torelli N (2015) Clustering of time series via non-parametric tail dependence estimation. Stat Pap 56(3):701–721
8. D'Urso P (2015) Fuzzy clustering. In: Meila M, Murtagh F, Rocci R (eds) Handbook of Cluster Analysis. Hennig C. Chapman & Hall,
9. Haerdle W, Nasekin S, Chuen D, Fai P (2014) TEDAS—Tail Event Driven ASset Allocation. Sfb 649 discussion papers, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2014-032.pdf

10. Haerdle W, Chuen D, Nasekin S, Ni X, Petukhina A (2015) Tail event driven ASset allocation: evidence from equity and mutual funds' markets. Sfb 649 discussion papers, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2015-045.pdf
11. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, data mining, inference, and prediction, Springer Series in Statistics, 2nd edn. Springer, New York
12. Kaufman L, Rousseeuw P (1990) Finding groups in data. Applied probability and statistics. Wiley Series in probability and mathematical statistics. John Wiley & Sons Inc., New York
13. Mantegna R (1999) Hierarchical structure in financial markets. Euro Phys J B 11(1):193–197
14. Patton AJ (2013) Copula methods for forecasting multivariate time series. In: Elliott G, Timmermann A (eds) Handbook of economic forecasting, vol 2. Elsevier, Oxford, pp 899–960
15. Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. J Risk 2:21–41

# An Upper Bound Estimation About the Sample Average of Interval-Valued Random Sets

**Xia Wang and Li Guan**

**Abstract** In this paper, we give an upper bound estimation about the probability of the event that the sample average of i.i.d. interval-valued random sets is included in a closed set. The main tool is Cramér theorem in the classic theory of large deviation principle about real-valued random variables.

## 1 Introduction

As we know, the theory of large deviation principle (LDP) deals with the asymptotic estimation of probabilities of rare events and provides exponential bound on probability of such events. In 1999, Cerf [1] proved LDP for sums of i.i.d. compact random sets in a separable type $p$ Banach space with respect to the Hausdorff distance $d_H$, which is called Cramér type LDP. For the Cramér type LDP, it considered the probability of the event that the sample average of i.i.d. random variables belongs to a closed set and an open set. However, in our paper, we consider the probability of the event that the sample average of i.i.d. interval-valued random sets is included in a closed set in $\mathbb{R}$, and give an upper bound estimation. The main tool is Cramér theorem in the classic theory of large deviation principle about real-valued random variables. Finally, we give an example about our main result.

The paper is structured as follows. Section 2 will give some preliminaries about interval-valued random sets. Our main results and proofs will be made in Sect. 3, and we give an example.

X. Wang (✉) · L. Guan
College of Applied Sciences, Beijing University of Technology, Beijing, China
e-mail: xiaking2008@163.com

L. Guan
e-mail: guanli@bjut.edu.cn

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_64

## 2 Preliminaries

Throughout our paper, we assume that $(\Omega, \mathcal{A}, P)$ is a complete probability space. Let $\mathcal{K}_{kc}(\mathbb{R})$ be the family of all non-empty compact convex subsets of $\mathbb{R}$, in fact, the element of non-empty compact convex subsets of $\mathcal{K}_{kc}(\mathbb{R})$ has the following form: $[a, b]$, $a < b$. Let $A = [A_1, A_2]$ and $B = [B_1, B_2]$ be two non-empty compact convex subsets of $\mathbb{R}$ and let $\lambda \in \mathbb{R}^+$, we can define addition and scalar multiplication by

$$A + B = \{a + b : \ a \in A, \ b \in B\} = [A_1 + B_1, A_2 + B_2],$$
$$\lambda A = \{\lambda a : \ a \in A\} = [\lambda A_1, \lambda A_2].$$

The Hausdorff distance on $\mathcal{K}_{kc}(\mathbb{R})$ is defined by

$$d_H(A, B) = \max\left\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\right\},$$

where $d(a, B) = \inf_{b\in B} |a - b|$.

Now we give the definition of interval-valued random sets. An interval-valued random set is a measurable mapping from the probability space $(\Omega, \mathcal{A}, P)$ to the space $(\mathcal{K}_{kc}(\mathbb{R}), \mathfrak{B}(\mathcal{K}_{kc}(\mathbb{R})))$ where $\mathfrak{B}(\mathcal{K}_{kc}(\mathbb{R}))$ is the Borel $\sigma$-field of $\mathcal{K}_{kc}(\mathbb{R})$ generated by the Hausdorff distance $d_H$. In fact, An interval-valued random set $X$ has the following form: $[X^{(l)}, X^{(r)}]$, where $X^{(l)}, X^{(r)}$ are real random variables and $X^{(l)} \le X^{(r)}$.

## 3 Main Results

Before we give our main theorem, we first show the following proposition.

**Proposition 1** *Let* $X_1 = [X_1^{(l)}, X_1^{(r)}], X_2 = [X_2^{(l)}, X_2^{(r)}], \cdots, X_n = [X_n^{(l)}, X_n^{(r)}]$ *be i.i.d. interval-valued random sets. Then* $X_1^{(l)}, X_2^{(l)}, \cdots, X_n^{(l)}$ *are i.i.d. random variables and* $X_1^{(r)}, X_2^{(r)}, \cdots, X_n^{(r)}$ *are also i.i.d. random variables.*

*Remark* In fact, the result in this Proposition is already known in set-valued set theory, but here, we still want to give a most elementary proof method for this Proposition.

*Proof* Here we only need to prove that $X_1^{(l)}, X_2^{(l)}, \cdots, X_n^{(l)}$ are i.i.d. random variables. For this aim, it is enough to prove that the following equations hold for any $x, y \in \mathbb{R}$,

$$P\{X_1^{(l)} \le x, X_2^{(l)} \le y\} = P\{X_1^{(l)} \le x\}P\{X_2^{(l)} \le y\}; \tag{1}$$
$$P\{X_1^{(l)} \le x\} = P\{X_2^{(l)} \le x\}. \tag{2}$$

In view of conditions given in this proposition: $X_1 = [X_1^{(l)}, X_1^{(r)}], X_2 = [X_2^{(l)}, X_2^{(r)}]$ are i.i.d. interval-valued random sets, so we have

$$P\{X_1 \in A, X_2 \in B\} = P\{X_1 \in A\}P\{X_2 \in B\},$$
$$P\{X_1 \in A\} = P\{X_2 \in A\},$$

for any closed set $A$ and $B$ of the space $(\mathcal{K}_{kc}(\mathbb{R}), d_H)$.

Now we take $A = \{[a, b] : a \leq x, b \geq a\}$, $B = \{[a, b] : a \leq y, b \geq a\}$. One hand, the sets $A, B$ are closed sets in the space $(\mathcal{K}_{kc}(\mathbb{R}), d_H)$. Here, we only prove $A$ is a closed set of the space $(\mathcal{K}_{kc}(\mathbb{R}), d_H)$. In fact, for any $x_0 \in A$, let $x_0 = [a_0, b_0]$, where $a_0 \leq x, b_0 \geq a_0$. In $\mathbb{R}$, there exists $a_n \in \mathbb{R}$, such that $a_n \leq x$ and $a_n \nearrow a_0$. Then $[a_n, b_0] \in A$ and $d_H([a_n, b_0], [a_0, b_0]) = a_0 - a_n \to 0$, as $n \to \infty$. So $A$ is a closed set of the space $(\mathcal{K}_{kc}(\mathbb{R}), d_H)$. On the other hand, $\{X_1 \in A\} = \{X_1^{(l)} \leq x\}$, $\{X_2 \in B\} = \{X_2^{(l)} \leq y\}$, $\{X_2 \in A\} = \{X_2^{(l)} \leq x\}$, then (1) and (2) have been proved.

Now, we present our main theorem.

**Theorem 1** *Let $X_1 = [X_1^{(l)}, X_1^{(r)}]$, $X_2 = [X_2^{(l)}, X_2^{(r)}]$, $\cdots$, $X_n = [X_n^{(l)}, X_n^{(r)}]$ be i.i.d. interval-valued random sets satisfying $Ee^{\lambda^{(l)}|X_1^{(l)}|} < \infty$ for some $\lambda^{(l)} > 0$, and $Ee^{\lambda^{(r)}|X_1^{(r)}|} < \infty$ for some $\lambda^{(r)} > 0$. Then for any closed set $C \subset \mathbb{R}$, we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{\frac{1}{n}\sum_{i=1}^{n} X_i \subset C\} \leq -(\inf_{x\in C} I^{(l)}(x) \vee \inf_{x\in C} I^{(r)}(x)), \tag{3}$$

*where*

$$I^{(l)}(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \log Ee^{\lambda X_1^{(l)}}\},$$
$$I^{(r)}(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \log Ee^{\lambda X_1^{(r)}}\}.$$

*Proof*

$$P\{\frac{1}{n}\sum_{i=1}^{n} X_i \subset C\} = P\{[\frac{1}{n}\sum_{i=1}^{n} X_i^{(l)}, \frac{1}{n}\sum_{i=1}^{n} X_i^{(r)}] \subset C\}$$
$$= P\{\frac{1}{n}\sum_{i=1}^{n} X_i^{(l)} \in C, \frac{1}{n}\sum_{i=1}^{n} X_i^{(r)} \in C\}$$
$$\leq P\{\frac{1}{n}\sum_{i=1}^{n} X_i^{(l)} \in C\} \wedge P\{\frac{1}{n}\sum_{i=1}^{n} X_i^{(r)} \in C\}.$$

Then

$$\limsup_{n\to\infty}\frac{1}{n}\log P\{\frac{1}{n}\sum_{i=1}^{n}X_i\subset C\}$$
$$\leq \limsup_{n\to\infty}\frac{1}{n}\log P\{\frac{1}{n}\sum_{i=1}^{n}X_i^{(l)}\in C\}\wedge \limsup_{n\to\infty}\frac{1}{n}\log P\{\frac{1}{n}\sum_{i=1}^{n}X_i^{(r)}\in C\}. \quad (4)$$

By Proposition 1, we know $X_1^{(l)}, X_2^{(l)}, \cdots, X_n^{(l)}$ are i.i.d. random variables and $X_1^{(r)}, X_2^{(r)}, \cdots, X_n^{(r)}$ are i.i.d. random variables, and since these random variables satisfy $Ee^{\lambda^{(l)}|X_1^{(l)}|}<\infty$ for some $\lambda^{(l)}>0$, and $Ee^{\lambda^{(r)}|X_1^{(r)}|}<\infty$ for some $\lambda^{(r)}>0$, then by the following Cramér theorem of classic theory of large deviation principles (see cf. [2, p.27 Theorem 2.2.3] or Appendix of this paper), then we have

$$\limsup_{n\to\infty}\frac{1}{n}\log P\{\frac{1}{n}\sum_{i=1}^{n}X_i^{(l)}\in C\}\leq -\inf_{x\in C}I^{(l)}(x), \quad (5)$$
$$\limsup_{n\to\infty}\frac{1}{n}\log P\{\frac{1}{n}\sum_{i=1}^{n}X_i^{(r)}\in C\}\leq -\inf_{x\in C}I^{(r)}(x). \quad (6)$$

Then combine (4), (5) and (6), we obtain (3).

**Corollary 1** *Let* $X_1=[X_1^{(l)}, X_1^{(r)}]$, $X_2$=$[X_2^{(l)}, X_2^{(r)}]$, $\cdots$, $X_n=[X_n^{(l)}, X_n^{(r)}]$ *be i.i.d. interval-valued random sets satisfying* $Ee^{\lambda^{(l)}|X_1^{(l)}|}<\infty$ *for some* $\lambda^{(l)}>0$, *and* $Ee^{\lambda^{(r)}|X_1^{(r)}|}<\infty$ *for some* $\lambda^{(r)}>0$. *Then for any* $a, b\in\mathbb{R}$, $a\leq b$, *we have*

$$\limsup_{n\to\infty}\frac{1}{n}\log P\{\frac{1}{n}\sum_{i=1}^{n}X_i\subset[a,b]\}\leq -(\inf_{x\in[a,b]}I^{(l)}(x)\vee\inf_{x\in[a,b]}I^{(r)}(x)),$$

*where*

$$I^{(l)}(x)=\sup_{\lambda\in\mathbb{R}}\{\lambda x-\log Ee^{\lambda X_1^{(l)}}\},$$
$$I^{(r)}(x)=\sup_{\lambda\in\mathbb{R}}\{\lambda x-\log Ee^{\lambda X_1^{(r)}}\}.$$

*Remark* From this Corollary, we know, under the case of real-valued random variables, this result is coherent with the Cramér theorem in the classic theory of large deviation principle about real-valued random variables.

**Corollary 2** *Let* $X_1=X_1^{(l)}=X_1^{(r)}$, $X_2=X_2^{(l)}=X_2^{(r)}$, $\cdots$, $X_n=X_n^{(l)}=X_n^{(r)}$ *and satisfy* $Ee^{\lambda^{(r)}|X_1^{(r)}|}<\infty$ *for some* $\lambda^{(r)}>0$. *Then for any closed set* $C\subset\mathbb{R}$, *we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{\frac{1}{n}\sum_{i=1}^{n} X_i \in C\} \leq - \inf_{x\in C} I^{(l)}(x),$$

*where*

$$I^{(l)}(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \log Ee^{\lambda X_1^{(l)}}\}.$$

**Corollary 3** *Let* $X_1 = X_1^{(l)} = X_1^{(r)}, X_2 = X_2^{(l)} = X_2^{(r)}, \cdots, X_n = X_n^{(l)} = X_n^{(r)}$ *and satisfy* $Ee^{\lambda^{(r)}|X_1^{(r)}|} < \infty$ *for some* $\lambda^{(r)} > 0$*. Then for any* $a, b \in \mathbb{R}, a \leq b$*, we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log P(\frac{1}{n}\sum_{i=1}^{n} X_i \in [a, b]) \leq - \inf_{x\in[a,b]} I^{(l)}(x),$$

*where*

$$I^{(l)}(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \log Ee^{\lambda X_1^{(l)}}\}.$$

*Example* In Theorem 1, let $X_1^{(l)}, X_2^{(l)}, \cdots, X_n^{(l)}$ be i.i.d random variables with the exponential distribution with parameter 1 and $X_i^{(r)} = X_i^{(l)} + \frac{1}{2}, X_i = [X_i^{(l)}, X_i^{(r)}]$, $i = 1, 2, \cdots, n$. Then we can get

$$I^{(l)}(x) = x - 1 - \log x, x > 0;\ I^{(r)}(x) = x - \frac{3}{2} - \log(x - \frac{1}{2}), x > \frac{1}{2}.$$

Now take $C = [2, 3]$, so $\inf_{2\leq x\leq 3} I^{(l)}(x) = 1 - \ln 2 \approx 0.69897$, $\inf_{2\leq x\leq 3} I^{(r)}(x) = 0.5 - \log 1.5 \approx 0.094535$, $\inf_{2\leq x\leq 3} I^{(l)}(x) \wedge \inf_{2\leq x\leq 3} I^{(r)}(x) = 0.5 - 5\log 1.5 \approx 0.094535$, then

$$\limsup_{n\to\infty} \frac{1}{n} \log P(\frac{1}{n}\sum_{i=1}^{n} X_i \subset [2, 3]) \leq -0.094535.$$

**Acknowledgments** Authors first thank the anonymous reviewers for those comments and suggestions, and then would like to thank the support from National Natural Science Foundation of China (No. 11401016, 11301015, 11571024) and Collaborative Innovation Center on Capital Social Construction and Social Management.

## Appendix

**Cramér theorem:** Let $X_1, X_2, \cdots$, be i.i.d random variables and satisfy $Ee^{\lambda|X_1|} < \infty$ for some $\lambda > 0$. Then for any closed set $F \subset \mathbb{R}$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{\frac{1}{n}\sum_{i=1}^{n} X_i \in F\} \le -\inf_{x\in F} I(x),$$

and for any open set $G \subset \mathbb{R}$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{\frac{1}{n}\sum_{i=1}^{n} X_i \in G\} \ge -\inf_{x\in G} I(x),$$

where

$$I(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \log Ee^{\lambda X_1}\}.$$

## References

1. Cerf R (1999) Large deviations for sums of i.i.d. random compact sets. Proc Am Math Soc 127:2431–2436
2. Dembo A, Zeitouni O (1998) Large deviations techniques and applications. 2nd edn. Springer
3. Li S, Ogura Y, Kreinovich V (2002) Limit Theorems and applications of set-valued and fuzzy-valued random variables. Kluwer Academic Publishers
4. Ogura Y, Li S, Xia Wang (2010) Large and moderate deviations of random upper semicontinuous functions. Stoch Anal Appl 28:350–376
5. Teran P (2005) A large deviation principle for random upper semicontimuous functions. Proc Am Math Soc 134:571–580
6. Teran P (2006) On Borel measurability and large deviations for fuzzy random variables. Fuzzy Sets Syst 157:2558–2568
7. Wang X (2013) Large deviations and moderate deviations for random sets and random upper semicontinuous functions. Int J Approximate Reasoning 54:378–392

# On Asymptotic Properties of the Multiple Fuzzy Least Squares Estimator

**Jin Hee Yoon, Seung Hoe Choi and Przemyslaw Grzegorzewski**

**Abstract** The multiple fuzzy linear regression model with fuzzy input–fuzzy output is considered. Assuming that fuzzy inputs and fuzzy outputs are modeled by triangular fuzzy numbers, we prove the consistency and asymptotic normality of the least squares estimators.

## 1 Introduction

The least squares method is the most widely used statistical technique to find unknown parameters of a regression model. However, there are many situations where observations for regression model cannot be described accurately. To record such data we need some approach to handle the uncertainty. Zadeh [31] introduced the concept of fuzzy sets to model imprecision or vagueness. Then Tanaka et al. [26] considered fuzzy regression analysis. Diamond [6] introduced fuzzy least squares estimation for triangular fuzzy numbers. Some authors have discussed the situation where both input and output are fuzzy [1, 2, 6, 13–17, 19, 20, 25, 27]. The others have studied the fuzzy model with crisp parameters [1, 6, 8, 16, 17, 19]. For situations where data has an error structure which is assumed in model, Diamond [7] and Näther [18, 21–23] introduced the fuzzy best linear unbiased estimators (FBLUEs). Kim et al. [17] established the asymptotic properties of fuzzy least squares estimators (FLSEs) in the case of a simple fuzzy linear regression model. Due to the complexity of expression of the least squares estimators some authors use $\alpha$-level sets to

J.H. Yoon
School of Mathematics and Statistics, Sejong University, Seoul, South Korea
e-mail: jin9135@sejong.ac.kr

S.H. Choi
School of Liberal Arts and Science, Korea Aerospace University, Goyang, South Korea
e-mail: shchoi@kau.ac.kr

P. Grzegorzewski (✉)
Faculty of Mathematics and Information Science, Systems Research Institute,
Polish Academy of Sciences, Warsaw University of Technology, Warsaw, Poland
e-mail: pgrzeg@ibspan.waw.pl

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances
in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_65

express the estimators [24, 25], while others separate the estimators into three parts: the mode and two spreads [4, 5, 18, 30]. Moreover, some authors do not express the formulas for the desired estimators but they found the estimates directly from the normal equations [6, 19]. To overcome these problems the *triangular fuzzy matrix* and suitable operations were defined in our previous studies [28, 29]. In this contribution we continue the examination of the fuzzy least squares estimator obtained there, focusing on its asymptotic properties.

The paper is organized as follows: in Sect. 2 we introduce basic notation and recall some facts used later in the contribution. In Sect. 3 we discuss the regression model proposed by Yoon and Choi [28, 29]. Then, in Sect. 3 we prove the asymptotic properties of the fuzzy least squares estimator dedicated to fuzzy inputs and fuzzy outputs modeled by triangular fuzzy numbers.

## 2 Preliminaries

Let $\mathcal{F}_T$ denote a family of all triangular fuzzy numbers. Each $A \in \mathcal{F}_T$ can be represented by an ordered triple, i.e. $A = (l_a, a, r_a)$, where $a$ is the mode of $A$, while $l_a$ and $r_a$ denote the lower and the upper bound of the support of $A$, respectively. Besides well-known basic operations on fuzzy numbers some other operations defined in $\mathcal{F}_T$ are sometimes useful. Let us recall here a few concepts proposed in [28]. From now on let us assume that $\mathcal{F}_T$ denote a family of all triangular fuzzy numbers defined on the non-negative real numbers $\mathbb{R}^*$.

**Definition 1** Let $X = (l_x, x, r_x)$ and $Y = (l_y, y, r_y)$ be any triangular fuzzy numbers. Then

$$X \diamond Y = l_x l_y + xy + r_x r_y, \tag{1}$$

$$X \otimes Y = (l_x l_y,\ xy,\ r_x r_y). \tag{2}$$

Clearly, the output of (1) is a real number, while the output of (2) belongs to $\mathcal{F}_T$. Based on *Zadeh's extension principle* [31] it is known that $\langle m_1,\ l_1,\ r_1\rangle_{LR} \oplus \langle m_2,\ l_2,\ r_2\rangle = \langle m_1 + m_2,\ l_1 + l_2,\ r_1 + r_2\rangle$, and

$$\lambda\langle m,\ l,\ r\rangle_{LR} = \begin{cases} \langle \lambda m, \quad \lambda l, \quad \lambda r\rangle_{LR} & \text{if} \quad \lambda > 0, \\ \langle \lambda m, -\lambda r, -\lambda l\rangle_{LR} & \text{if} \quad \lambda < 0, \\ \langle 0, 0, 0\rangle_{LR} & \text{if} \quad \lambda = 0. \end{cases}$$

Further on we'll also need some operations defined on the matrices.

**Definition 2** A triangular fuzzy matrix (t.f.m.) is a matrix whose elements are triangular fuzzy numbers. For given two $n \times n$ triangular fuzzy matrices $\tilde{\Gamma} = [X_{ij}]$ and $\tilde{\Lambda} = [Y_{ij}]$ their addition $\tilde{\Gamma} \oplus \tilde{\Lambda}$ is defined by the $n \times n$ t.f.m. $\tilde{\Sigma} = [Z_{ij}]$, where $Z_{ij} = X_{ij} \oplus Y_{ij}$. Moreover, two products $\tilde{\Gamma} \diamond \tilde{\Lambda}$ and $\tilde{\Gamma} \oplus \tilde{\Lambda}$, the product of crisp matrix $A = [a_{ij}]$ and t.f.m. $\tilde{\Gamma}$ and the scalar multiplication $k\tilde{\Gamma}$, where $k \in \mathbb{R}$, are defined as follows

$$\tilde{\Gamma} \diamond \tilde{\Lambda} = [\sum_{k=1}^{n} X_{ik} \diamond Y_{kj}], \qquad \tilde{\Gamma} \otimes \tilde{\Lambda} = [\bigoplus_{k=1}^{n} X_{ik} \otimes Y_{kj}],$$

$$\tilde{A}\tilde{\Gamma} = [\bigoplus_{k=1}^{n} a_{ik} X_{kj}], \qquad k\tilde{\Gamma} = [kX_{ij}].$$

We denote by $M_{\mathbb{R}^*}$ the set of all $n \times n$ real crisp matrices with nonnegative elements and let $M_{\mathcal{F}_\mathcal{T}}$ be the set of all fuzzy element matrices on $\mathcal{F}_\mathcal{T}$. Of course, $\tilde{\Gamma} \oplus \tilde{\Lambda}$, $\tilde{\Gamma} \otimes \tilde{\Lambda}, \tilde{A}\tilde{\Gamma} \in M_{\mathcal{F}_\mathcal{T}}$ and $\tilde{\Gamma} \diamond \tilde{\Lambda} \in M_{\mathbb{R}^*}$. We can also define the following three types of fuzzy scalar multiplications of a crisp matrix.

**Definition 3** For given $X \in \mathcal{F}_\mathcal{T}$, $\tilde{A} = [a_{ij}] \in M_{\mathbb{R}^*}$ and $\tilde{\Gamma} = [X_{ij}] \in M_{\mathcal{F}_\mathcal{T}}$, we define three fuzzy scalar multiplications, $XA$, $X \diamond \tilde{\Gamma}$ and $X \otimes \tilde{\Gamma}$, where

$$X\tilde{A} = [a_{ij}X], \quad X \diamond \tilde{\Gamma} = [X \diamond X_{ij}], \quad X \otimes \tilde{\Gamma} = [X \otimes X_{ij}]. \tag{3}$$

Finally, we will consider the convergence defined as follows.

**Definition 4** For $X = (l_x, x, r_x),\ Y = (l_y, y, r_y) \in \mathcal{F}_\mathcal{T}$ we say that $X \longrightarrow Y$ if $l_x \rightarrow l_y,\ x \rightarrow y$ and $r_x \rightarrow r_y$. Moreover, for $\tilde{\Gamma} = [X_{ij}],\ \tilde{\Lambda} = [Y_{ij}] \in M_{\mathcal{F}_\mathcal{T}}$ we say that $\tilde{\Gamma} \longrightarrow \tilde{\Lambda}$ if $X_{ij} \rightarrow Y_{ij}$ for all $i, j = 1, \ldots, n$.

We end this section by citing some theorems concerning the Central Limit Theorem (CLT) and Strong Law of Large Numbers (SLLN) for martingales [10] which will be useful in the proof of the main result in this contribution.

**Theorem 1** (Hajék-Sidăk CLT) *Let $\{X_n\}$ be a sequence of i.i.d. random variables (r.v.'s) with mean $\mu$ and finite variance $\sigma^2$. Let $\{\boldsymbol{c}_n\}$ be a sequence of real vectors $\boldsymbol{c}_n = (c_{n1}, \cdots, c_{nn})^t$. If*

$$\left(\max_{1\le i\le n} c_{ni}^2\right)\left(\sum_{i=1}^{n} c_{ni}^2\right)^{-1} \longrightarrow 0 \quad \text{as } n \rightarrow \infty,$$

*then*

$$Z_n = \frac{\sum_{i=1}^{n} c_{ni}(X_i - \mu)}{\sigma^2 \sum_{i=1}^{n} c_{ni}^2} \xrightarrow{L} N(0, 1),$$

*where the notation $\xrightarrow{L}$ stands for convergence in law.*

**Theorem 2** (SLLN for martingales) *Let $S_n = \sum_{i=1}^{n} X_i$, $n \geq 1$, be a martingale such that $E|X_k|^p < \infty$ for $k \geq 1$ and $1 \leq p \leq 2$. Suppose that $\{b_n\}$ is a sequence of positive constants increasing to $\infty$ as $n \rightarrow \infty$, and $\sum_{i=1}^{n} E[X_i^2]/b_i^2 < \infty$. Then $S_n/b_n \xrightarrow{a.s.} 0$, where the notation $\xrightarrow{a.s.}$ means converges almost surely.*

**Theorem 3** (Courant-Fisher minimax theorem) *For any $n \times n$ real symmetric matrix $A$ its eigenvalues $\lambda_1 \leqslant \ldots \leqslant \lambda_n$ satisfy*

$$\lambda_k = \min_{dim(C)=k} \max_{||x||=1, x \in C} < Ax, x >, \tag{4}$$

*where $C$ is a subspace of $\mathbb{R}^n$.*

## 3 Fuzzy Least Squares Estimation

Throughout this paper we consider the following linear regression model with fuzzy inputs and fuzzy outputs

$$Y_i = \beta_0 \oplus \beta_1 X_{i1} \oplus \ldots \oplus \beta_p X_{ip} \oplus \Phi_i, \quad i = 1, \ldots, n, \tag{5}$$

where $X_{ij} = (l_{x_{ij}}, x_{ij}, l_{x_{ij}})$ and $Y_i = (ly_i, y_i, r_{y_i})$,$j = 1, \ldots, p$, while $\beta_j$ denote unknown crisp regression parameters to be estimated from the observations of $Y_i$ and $X_{ij}$. Moreover, let $\Phi_i$, $i = 1, \ldots, n$, denote fuzzy error terms which express both randomness and fuzziness allowing negative spreads [3, 8, 17], i.e. $\Phi_i = (\theta_i^l, \epsilon_i, \theta_i^r)$, where $\theta_i^l, \epsilon_i, \theta_i^r$ are crisp random variables. We suggest the following assumptions to be satisfied by these random variables.

**Assumption A**

(A1) $\epsilon_i$ are i.i.d. r.v.'s such that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma_\epsilon^2 < \infty$.
(A2) $\theta_i^r$, $\theta_i^l$ are i.i.d. r.v.'s such that $E(\theta_i^r) = 0$, $E(\theta_i^l) = 0$, $Var(\theta_i^r) = \sigma_r^2 < \infty$ and $Var(\theta_i^l) = \sigma_l^2 < \infty$.
(A3) $\epsilon_i$, $\theta_i^r$ and $\theta_i^l$ are mutually uncorrelated.

Defining the following design matrix $\tilde{X} = [(l_{x_{ik}}, x_{ik}, r_{x_{ik}})]_{n \times (p+1)}$, where

$$\tilde{X} = \begin{bmatrix} (1, 1, 1) & (l_{x_{11}}, x_{11}, r_{x_{11}}) & \cdots & (l_{x_{1p}}, x_{1p}, r_{x_{1p}}) \\ \vdots & \vdots & \ddots & \vdots \\ (1, 1, 1) & (l_{x_{n1}}, x_{n1}, r_{x_{n1}}) & \cdots & (l_{x_{np}}, x_{np}, r_{x_{np}}) \end{bmatrix},$$

and a vector $\tilde{\mathbf{y}} = [(l_{y_i}, y_i, r_{y_i})]_{n \times 1} = [(l_{y_1}, y_1, r_{y_1}), \cdots, (l_{y_n}, y_n, r_{y_n})]^t$ and assuming that $det(\tilde{X}^t \diamond \tilde{X}) \neq 0$, by [28] we obtain the following least squares estimator

$$\hat{\boldsymbol{\beta}} = (\tilde{X}^t \diamond \tilde{X})^{-1} \tilde{X}^t \diamond \tilde{\mathbf{y}}. \tag{6}$$

The following facts will be later useful.

**Lemma 1** (see [29]) *Let $\tilde{\Gamma}_{m\times n} \in M_{\mathcal{F}_\mathcal{T}}$ and $\tilde{\mathbf{y}}_{n\times 1} = [Y_j] = [(l_{y_j}, y_j, r_{y_j})] \in M_{\mathcal{F}_\mathcal{T}}$, where $j = 1, \dots, n$. Then*

$$\begin{aligned} Var(\tilde{\Gamma} \diamond \tilde{\mathbf{y}}) &= (\tilde{\Gamma} \otimes \tilde{\Sigma}) \diamond \tilde{\Gamma}^t \\ &= \sigma_{\Phi}^2 \diamond (\tilde{\Gamma} \otimes \tilde{\Gamma}^t). \end{aligned} \tag{7}$$

*Assuming $Var(l_{y_j}) = \sigma^2_{l_{y_j}}$, $Var(y_j) = \sigma^2_{y_j}$ and $Var(r_{y_j}) = \sigma^2_{r_{y_j}}$, the matrix $\tilde{\Sigma}_{n\times n} \in M_{\mathcal{F}_\mathcal{T}}$ has diagonals $\tilde{\Sigma}_{jj} = (\sigma^2_{l_{y_j}}, \sigma^2_{y_j}, \sigma^2_{r_{y_j}}) = (\sigma_l^2, \sigma_\epsilon^2, \sigma_r^2) = \sigma^2_\Phi$, while $\tilde{\Sigma}_{jl} =$ $(0, 0, 0)$ for $j, l = 1, \dots, n$ and $j \neq l$.*

**Theorem 4** (see [29]) *Let $\hat{\boldsymbol{\beta}}$ be the least squares estimator* (6). *Then*

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2_\Phi \diamond \left( (\tilde{X}^t \diamond \tilde{X})^{-1} (\tilde{X}^t \otimes \tilde{X}) (\tilde{X}^t \diamond \tilde{X})^{-1} \right). \tag{8}$$

## 4 Asymptotic Properties

To prove asymptotic properties of our estimator the following additional assumption are required besides Assumption A given in Sect. 4.

**Assumption B.**

(B1) $\max\limits_{1\le i\le n} \left( \tilde{\mathbf{x}}_i{}^t (X^t \diamond \tilde{X})^{-1} \right) \diamond \tilde{\mathbf{x}}_i \to 0$ as $n \to \infty$, where $\tilde{\mathbf{x}}_i{}^t$ denotes the $i$-th row of $\tilde{X}$.

(B2) $n(\tilde{X}^t \diamond \tilde{X})^{-1} (\tilde{X}^t \otimes \tilde{X}) (\tilde{X}^t \diamond \tilde{X})^{-1} \to \tilde{\Pi}$ as $n \to \infty$ for some $\tilde{\Pi} \in M_{\mathcal{F}_\mathcal{T}}$.

Now we are able to formulate the main result of this contribution.

**Theorem 5** *If model* (5) *satisfies Assumption A and Assumption B then the least squares estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal, i.e.*

$$\sqrt{n} \left( \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \right) \xrightarrow{L} N_{p+1} \left( \mathbf{0},\ \sigma^2_\Phi \diamond \tilde{\Pi} \right), \tag{9}$$

*where $\sigma^2_\Phi = (\sigma_l^2, \sigma_\epsilon^2, \sigma_r^2)$.*

*Proof* By (6) one can find that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n &= (\tilde{X}^t \diamond \tilde{X})^{-1} (\tilde{X}^t \diamond \tilde{\mathbf{y}}) = \left( (\tilde{X}^t \diamond \tilde{X})^{-1} \tilde{X}^t \right) \diamond \tilde{\mathbf{y}} \\ &= (\tilde{X}^t \diamond \tilde{X})^{-1} \tilde{X}^t \diamond (\tilde{X}\boldsymbol{\beta} + \Phi) = \boldsymbol{\beta} + (\tilde{X}^t \diamond \tilde{X})^{-1} (\tilde{X}^t \diamond \Phi) \\ &= \boldsymbol{\beta} + \left( (\tilde{X}^t \diamond \tilde{X})^{-1} \tilde{X}^t \right) \diamond \Phi, \end{aligned}$$

so, consequently, $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \left( (\tilde{X}^t \diamond \tilde{X})^{-1} \tilde{X}^t \right) \diamond \Phi$.

Let $\boldsymbol{\lambda}_n \in R^{p+1}(\boldsymbol{\lambda}_n \neq \mathbf{0})$ be an arbitrary but fixed vector. Moreover, let $Z_n = {\boldsymbol{\lambda}_n}^t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \tilde{\boldsymbol{C}}_n{}^t \diamond \boldsymbol{\Phi} \in R$, where $\tilde{\boldsymbol{C}}_n = \tilde{X}(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n \in M_{\mathcal{F}_\mathcal{T}}$. If we denote $\tilde{\mathbf{C}}_n{}^t = [C_{n1}, \cdots, C_{nn}]$, where $C_{n1}, \cdots, C_{nn} \in \mathcal{F}_\mathcal{T}$, then by [28]

$$\begin{aligned}\sum_{i=1}^{n} C_{ni} \diamond C_{ni} = \tilde{\mathbf{C}}_n^t \diamond \tilde{\mathbf{C}}_n \\ &= \left(\boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}\tilde{X}^t\right) \diamond \left(\tilde{X}(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n\right) \\ &= \boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}(\tilde{X}^t \diamond \tilde{X})(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n \\ &= \boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n.\end{aligned}$$

We claim that $\tilde{\mathbf{C}}_n$ satisfies the regularity condition of Theorem 2. Then we can obtain the asymptotic distribution of $Z_n$.

Let $\tilde{\mathbf{x}}_i^t$ be the $i$th row of $\tilde{X}$. Then we get $C_{ni} = \mathbf{x}_i^t(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n$. Since $C_{ni} \in \mathcal{F}_\mathcal{T}$ we have $C_{ni}^t = C_{ni}$. Hence

$$\begin{aligned}C_{ni} \diamond C_{ni} &= C_{ni}^t \diamond C_{ni} \\ &= (\boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}\mathbf{x}_i) \diamond (\mathbf{x}_i^t(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n) \\ &= \boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}(\mathbf{x}_i \diamond \mathbf{x}_i^t)(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n.\end{aligned}$$

Therefore, by Theorem 3 and [28]

$$\sup_{\boldsymbol{\lambda}_n} \frac{\boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}(\tilde{\mathbf{x}}_i \diamond \tilde{\mathbf{x}}_i{}^t)(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n}{\boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n} \tag{10}$$

becomes

$$\mathrm{ch}_{max}[(\tilde{X}^t \diamond \tilde{X})^{-1}(\tilde{\mathbf{x}}_i \diamond \tilde{\mathbf{x}}_i{}^t)] = \left(\tilde{\mathbf{x}}_i{}^t(\tilde{X}^t \diamond \tilde{X})^{-1}\right) \diamond \tilde{\mathbf{x}}_i,$$

where $\mathrm{ch}_{max}(Q)$ stands for the largest characteristic value of matrix $Q$. Thus,

$$\begin{aligned}&\sup_{\boldsymbol{\lambda}_n} \max_i \frac{C_{ni} \diamond C_{ni}}{\sum_{i=1}^n C_{ni} \diamond C_{ni}} \\ &\quad= \max_i \sup_{\boldsymbol{\lambda}_n} \frac{\boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}(\tilde{\mathbf{x}}_i \diamond \tilde{\mathbf{x}}_i{}^t)(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n}{\boldsymbol{\lambda}_n^t(\tilde{X}^t \diamond \tilde{X})^{-1}\boldsymbol{\lambda}_n} \\ &\quad= \max_i \left(\tilde{\mathbf{x}}_i{}^t(\tilde{X}^t \diamond \tilde{X})^{-1}\right) \diamond \tilde{\mathbf{x}}_i,\end{aligned}$$

which, by assumption (B1), converges to 0 as $n \to \infty$. It means that

$$\max_{1 \le i \le n} \frac{C_{ni} \diamond C_{ni}}{\sum_{i=1}^n C_{ni} \diamond C_{ni}} \to 0$$

as $n \to \infty$. So, by Theorem 1, we obtain

$$Z_n = \boldsymbol{\lambda}_n{}^t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{L} N\,(0, 1)\,.$$

On the other hand one may notice that

$$\begin{aligned} Var\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\right) &= n\sigma_\Phi^2 \diamond \left((\tilde{X}^t \diamond \tilde{X})^{-1}(\tilde{X}^t \otimes \tilde{X})(\tilde{X}^t \diamond \tilde{X})^{-1}\right) \\ &\longrightarrow \sigma_\Phi^2 \diamond \tilde{\Pi}, \end{aligned}$$

as $n \to \infty$ (by assumption (B2)). Thus,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{L} N_{p+1}\left(\mathbf{0},\ \sigma_\Phi^2 \diamond \tilde{\Pi}\right),$$

which completes the proof. □

## 5 Conclusions

Asymptotic theory often makes it possible to carry out the analysis which cannot be obtained within a finite sample theory. In this paper we proved some asymptotic properties of estimators in the multiple fuzzy input-output regression model. We were focused especially on consistency and asymptotic normality of those estimators. To reach the goal we have introduced a suitable matrix, called *triangular fuzzy matrix*, and applied some operations provided in our previous studies (see [28, 29]).

Although the aforementioned asymptotic properties were discussed only for fuzzy inputs and outputs modeled by triangular fuzzy numbers, they could be easily extended to trapezoidal fuzzy numbers or *LR*-fuzzy numbers by defining adequate fuzzy matrices and corresponding operations. Moreover, after constructing suitable mathematical tools further research should be undertaken to examine the analogous results in another models, like the regression model with fuzzy parameters or crisp-input and fuzzy-output regression model.

## References

1. Bargiela A, Pedrycz A, Nakashima T (2007) Multiple regression with fuzzy data. Fuzzy Sets Syst 158:2169–2188
2. Celminš A (1987) Least squares model fitting to fuzzy vector data. Fuzzy Sets Syst 22:245–269
3. Chang P-T, Lee ES (1994) Fuzzy linear regression with spreads unrestricted in sign. Comput Math Appl 28:61–70
4. Chang P-T, Lee ES (1994) Fuzzy least absolute deviations regression and the conflicting trends in fuzzy parameters. Comput Math Appl 28:89–101

5. Choi SH, Yoon JH (2010) General fuzzy regression using least squares method. Int J Syst Sci 41:477–485
6. Diamond P (1988) Fuzzy least squares. Inform Sci 46:141–157
7. Diamond P (1989) Fuzzy kriging. Fuzzy Sets Syst 33:315–332
8. Diamond P, Körner R (1997) Extended fuzzy linear models and least squares estimates. Comput Math Appl 33:15–32
9. Diamond P, Kloeden P (1994) Metric spaces of fuzzy sets: theory and applications. World Scientific
10. Drygas H (1976) Weak and strong consistency of the least square estimates in regression models. Z Wahrscheinlickeitstheorie und Verw, Gebiete 34:119–127
11. D'Urso P, Gastaldi T (2000) A least-squares approach to fuzzy linear regression analysis. Comput Stat Data Anal 34:427–440
12. D'Urso P (2003) Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Comput Stat Data Anal 42:47–72
13. González-Rodríguez G, Blanco A, Colubi A, Lubiano MA (2009) Estimation of a simple linear regression model for fuzzy random variables. Fuzzy Sets Syst 160:357–370
14. Grzegorzewski P, Mrowka E (2003) Linear regression analysis for fuzzy data. In: Proceedings of the 10th IFSA World Congress—IFSA 2003, Istanbul, Turkey, 29 Jun–2 July 2003, pp 228–231
15. Grzegorzewski P, Mrowka E (2004) Regression analysis with fuzzy data. In: Grzegorzewski P, Krawczak M, Zadrony S (eds) Soft computing—tools, techniques and applications. Exit, Warszawa, pp 65–76
16. Hong DH, Hwang C (2004) Extended fuzzy regression models using regularization method. Inform Sci 164:31–46
17. Kim HK, Yoon JH, Li Y (2008) Asymptotic properties of least squares estimation with fuzzy observations. Inform Sci 178:439–451
18. Körner R, Näther W (1998) Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. Inform Sci 109:95–118
19. Ming M, Friedman M, Kandel A (1997) General fuzzy least squares. Fuzzy Sets Syst 88:107–118
20. Nasrabadi MM, Nasrabadi E (2004) A mathematical-programming approach to fuzzy linear regression analysis. Appl Math Comput 155:873–881
21. Näther W (1997) Linear statistical inference for random fuzzy data. Statistics 29:221–240
22. Näther W (2000) On random fuzzy variables of second order and their application to linear statistical inference with fuzzy data. Metrika 51:201–221
23. Näther W (2006) Regression with fuzzy random data. Comput Stat Data Anal 5:235–252
24. Parchami A (2006) M. Mashinchi, Fuzzy estimation for process capability indices. Inform Sci 177:1452–1462
25. Sakawa M, Yano H (1992) Multiobjective fuzzy linear regression analysis for fuzzy input-output data. Fuzzy Sets Syst 47:173–181
26. Tanaka H, Uejima S, Asai K (1982) Linear regression analysis with fuzzy model. IEEE Trans Syst Man Cybern 12:903–907
27. Yang M, Lin T (2002) Fuzzy least-squares linear regression analysis for fuzzy input-output data. Fuzzy Sets Syst 126:389–399
28. Yoon JH, Choi SH (2013) Fuzzy least squares estimation with new fuzzy operations. Adv Intell Syst Comput 190:193–202
29. Yoon JH, Jung HY, Lee WJ, Choi SH (2014) Optimal properties of a fuzzy least estimator based on new operations. In: Proceeding of SCIS&ISIS 2014, pp 36–41
30. Yoon JH, Choi SH (2007) Separate fuzzy regression with crisp input and fuzzy output. J Korean Data Inform Sci 18:301–314 (in Korean)
31. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353

# Author Index

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4