Advances in Intelligent Systems and Computing 315

Przemysław Grzegorzewski
Marek Gagolewski
Olgierd Hryniewicz
María Ángeles Gil *Editors*

Strengthening Links between Data Analysis and Soft Computing

Springer

Advances in Intelligent Systems and Computing

Volume 315

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

Przemysław Grzegorzewski · Marek Gagolewski
Olgierd Hryniewicz · María Ángeles Gil
Editors

Strengthening Links between Data Analysis and Soft Computing

Springer

Editors
Przemysław Grzegorzewski
Polish Academy of Sciences
Systems Research Institute
ul. Newelska 6, 01-447 Warsaw, Poland

and

Faculty of Mathematics and
Information Sciences
Warsaw University of Technology
ul. Koszykowa 75, 01-661 Warsaw, Poland

Marek Gagolewski
Polish Academy of Sciences
Systems Research Institute
ul. Newelska 6, 01-447 Warsaw, Poland

and

Faculty of Mathematics and
Information Sciences
Warsaw University of Technology
ul. Koszykowa 75, 01-661 Warsaw, Poland

Olgierd Hryniewicz
Polish Academy of Sciences
Systems Research Institute
Warsaw
Poland

María Ángeles Gil
Departamento de Estadistica
Universidad de Oviedo
Oviedo
Spain

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-10764-6 ISBN 978-3-319-10765-3 (eBook)
DOI 10.1007/978-3-319-10765-3

Library of Congress Control Number: 2014948458

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Probability and statistics were the only well-founded theories of uncertainty for a long time. However, during last forty years, in such areas like decision theory, artificial intelligence or information processing, numerous approaches extending or orthogonal to the existing theory of probability and mathematical statistics have been successfully developed. These new approaches have appeared, either on their own like fuzzy set theory, possibility theory, rough sets, or having their origin in probability theory itself, like imprecise probability, belief functions, fuzzy random variables.

The common feature of all those attempts is to allow for a more flexible modeling of imprecision, uncertainty, vagueness and ignorance. The proposed new methods are softer than the traditional theories and techniques because being less rigid they more easily adapt to the actual nature of information.

Wide range of applications still reveals the need for soft extensions of classical probability and statistical tools. For example, in data analysis and data mining it is becoming increasingly clear that integrating fuzzy sets and probability can lead to more robust and interpretable models which better capture all kinds of the information contained in data. Also, in science and engineering the need to analyze and model the true uncertainty associated with complex systems requires a more sophisticated representation of ignorance than that provided by uninformative Bayesian priors.

Several years ago, the need was felt to establish a recurrent forum for exchanging of ideas and discussing new trends that enlarge the statistical and uncertainty modeling traditions, towards a flexible and more specific handling of incomplete or subjective information. This idea resulted in a series of biannual international conferences on **Soft Methods in Probability and Statistics** (SMPS), organized for the first time in Warsaw in 2002. Subsequent events in this series took place in Oviedo (2004), Bristol (2006), Toulouse (2008), Oviedo/Mieres (2010) and Konstanz (2012).

This volume is a collection of papers presented at the 7th International Conference on Soft Methods in Probability and Statistics – SMPS 2014, held in Warsaw (Poland) on September 22–24, 2014. The conference was organized by the Polish Operational and Systems Research Society, Systems Research Institute of the Polish Academy of Sciences and the Faculty of Mathematics and Information Science of Warsaw University of Technology.

The volume contains three sections. The first one is dedicated to general aspects of uncertainty modeling and processing – starting from some fundamental problems, through various tools and approaches including Bayesian analysis, fuzzy sets and their generalizations, fuzzy equations etc. Part two encloses contributions devoted to soft methods in statistics emphasizing robust analysis and hypotheses testing in the presence of imprecise data. Part three consists of papers oriented on soft methods in data analysis. Here we find contributions concentrated on statistical methods in data mining, likewise combing statistical and soft computing perspectives – both more theoretical and oriented on applications.

The editors are grateful to contributing authors, invited speakers, Program Committee members and additional referees who made it possible to put together the attractive program of the conference. We 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.

Warsaw, June 23, 2014

Przemysław Grzegorzewski
Marek Gagolewski
Olgierd Hryniewicz
María Ángeles Gil

Organization

Executive Board

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

Program Committee

Bernard de Baets (Gent, Belgium)
Christian Borgelt (Mieres, Spain)
Giulianella Coletti (Perugia, Italy)
Ana Colubi (Oviedo, Spain)
Ines Couso (Oviedo, Spain)
Didier Dubois (Toulouse, France)
Fabrizio Durante (Bolzano, Italy)
Pierpaolo D'urso (Roma, Italy)
María Ángeles Gil (Oviedo, Spain)
Gil González Rodríguez (Oviedo, Spain)
Przemysław Grzegorzewski (Warsaw, Poland)
Olgierd Hryniewicz (Warsaw, Poland)
Eyke Hüllermeier (Marburg, Germany)
Piotr Jaworski (Warsaw, Poland)
Janusz Kacprzyk (Warsaw, Poland)
Frank Klawonn (Braunschweig/Wolfenbüttel, Germany)
Jacek Koronacki (Warsaw, Poland)
Rudolf Kruse (Magdeburg, Germany)
Mark Last (Negev, Israel)
Jonatan Lawry (Bristol, UK)
Radko Mesiar (Bratislava, Slovakia)
Enrique Miranda (Oviedo, Spain)
Susana Montes (Oviedo, Spain)

Zbigniew Nahorski (Warsaw, Poland)
Beloslav Riecan (Banská Bystrica, Slovakia)
Daniel Sanchez (Granada, Spain)
Martin Stepnicka (Ostrava, Czech Republic)
Seyed Mahmood Taheri (Isfahan, Iran)
Wolfgang Trutschnig (Salzburg, Austria)
Stefan Van Aelst (Ghent, Belgium)
Barbara Vantaggi (Roma, Italy)
Reinhard Viertl (Vienna, Austria)
Berlin Wu (Taipei, Taiwan)

Additional Reviewers

Adrian Ban, Humberto Bustince, Lucian Coroianu, Anna Czapkiewicz, Glad Deschrijver, Maria Brigida Ferraro, Marek Gagolewski, Piotr Nowak, Anna Olwert, Anna Stachowiak, Luciano Stefanini, Valentin Todorov, Jin Hee Yoon, Jinping Zhang

Organizing Committee

Anna Cena
Marek Gagolewski – publication and conference website chair
Przemysław Grzegorzewski – local chair
Adam Kołacz
Karol Opara
Barbara Żogała-Siudem

Contents

Part I: Soft Methods in Uncertainty Modeling and Processing

Models and Decisions

Interval-Valued and IF Sets

Bayesian Methods

Fuzzy Equations

Part II: Soft Methods in Statistics

Robust Statistics

Hypotheses Testing

Part III: Soft Methods in Data Analysis

Statistical Methods in Data Mining

Soft Computing and Statistics

Combining Soft and Stochastic Methodologies in Applications

Part I
Soft Methods in Uncertainty Modeling and Processing

Deciding under Ignorance: In Search of Meaningful Extensions of the Hurwicz Criterion to Decision Trees

Didier Dubois, Hélène Fargier, Romain Guillaume, and Caroline Thierry

IRIT, CNRS and Université de Toulouse, France
{dubois,fargier,guillaum,thierry}@irit.fr

Abstract. The major paradigm for sequential decision under uncertainty is expected utility. This approach has many good features that qualify it for posing and solving decision problems, especially dynamic consistency and computational efficiency via dynamic programming. However, when uncertainty is due to sheer lack of information, and expected utility is no longer a realistic criterion, the approach collapses because dynamic consistency becomes counterintuitive and the global non-expected utility criteria are no longer amenable to dynamic programming. In this paper we argue against Resolute Choice strategies, following the path opened by Jaffray, and suggest that the dynamic programming methodology may lead to more intuitive solutions respecting the Consequentialism axiom, while a global evaluation of strategies relying on lottery reduction is questionable.

1 Introduction

The traditional approach to multiple stage decision processes under the probabilistic approach [9] is based on decision trees. A decision tree is a graphical structure containing chance nodes and decision nodes. A strategy is the assignment of a decision (i.e. a chance node) to each decision node and each strategy turns the decision tree into a probability tree. This probability tree characterizes a unique probability distribution on the space of final states, and the (global) expected utility of the strategy over the state space can be computed. The optimal strategy is then the one with maximal expected utility.

This model has several features that make computations tractable [9]: Any substrategy of an optimal strategy is optimal with respect to the corresponding decision subtree, the optimal strategy can be computed by means of dynamic programming from the leaves to the root of the decision tree.

The appeal of this approach is also due to three properties that it verifies:

- *Dynamic Consistency*: When reaching a decision node by following an optimal strategy, the best decision at this node is the one that had been considered so when computing this strategy, i.e. prior to applying it.
- *Consequentialism*: the best decision at each step of the decision tree only depends on potential consequences at this point.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_1

- *Tree Reduction*: The result of the dynamic programming procedure on the decision tree comes down to optimizing the criteria defined on the state space via the probability distribution obtained from each strategy via the reduction of lotteries.

More recently, with the emergence of non-additive uncertainty theories, the decision tree approach has been adapted to new decision criteria that differ from expected utility [10], but generalize known but less reputed criteria such as Wald maximin criterion or Hurwicz criterion: for instance lower expected utility with respect to a set of priors [2] or Jaffray's belief function extension of Hurwicz criterion [6]. These criteria turn out to be incompatible with the three above assumptions in the sequential decision setting [4]. In particular, they violate Dynamic Consistency, and optimizing the non-expected utility criterion cannot be carried out using dynamic programming [3]. Some authors tend to privilege Dynamic Consistency and Tree Reduction and are ready to give up Consequentialism (e.g., the Resolute Choice approach [1]). Another approach called Veto-process has been proposed by Jaffray [7]. It insists on the fact that Resolute Choice is not acceptable since a normally behaved decision-maker is consequentialist.

The aim of this paper is to provide more arguments in favor of Consequentialism as a natural property to be preserved when uncertainty accounts for incomplete information rather than frequentist probability, while questioning Resolute Choice. We follow the line initiated by Jaffray [7] who introduced the so-called Veto-process in the frame of decision under total uncertainty. First, we present the background on decision trees under pure uncertainty and the Hurwicz criterion. Then, we illustrate Resolute Choice, showing its paradoxical behavior on a example. Then we present and discuss two alternatives to Resolute Choice, inspired by the Veto-process philosophy.

2 Background

In this section, we first recall the definition of the Hurwicz criterion and decision trees under uncertainty.

Consider first simple, non sequential decision problems under complete uncertainty: each decision δ is characterized by the multi set of consequences E_δ it can lead to - or equivalently a simple, non probabilitistic lottery. Given a utility function $(u(s))$ capturing the attractiveness of each of these consequences, a usual way to taking into account the optimism of the decision-maker under total uncertainty is to use the Hurwicz criterion [5]. The worth of a simple lottery δ is then:

$$H(\delta) = \alpha \times \min_{s \in E_\delta} u(s) + (1 - \alpha) \times \max_{s \in E_\delta} u(s).$$

where $\alpha \in [0, 1]$ is the degree of optimism.

When the decision problem is sequential and fully observable, we shall use decision trees [9] a graphical representations of the problem. This framework proposes an explicit modeling, representing each possible scenario by a path from the root to the leaves of the tree. Formally, the graphical component of a decision tree $\mathcal{T}$ is composed of a set of nodes $\mathcal{N}$ and a set of edges $\mathcal{E}$ such that the set $\mathcal{N}$ contains three kinds of nodes:

- $\mathcal{D} = \{d_0, \ldots, d_m\}$ is the set of decision nodes (represented by rectangles).
- $\mathcal{LN} = \{ln_1, \ldots, ln_k\}$ is the set of leaves, that represent final states in $\mathcal{S} = \{s_1, \ldots, s_k\}$; such states can be evaluated thanks to a utility function: $\forall s_i \in \mathcal{S}$, $u(s_i)$ is the utility of being eventually in state s_i (in node ln_i). For the sake of simplicity we assume that only leave nodes lead to utilities.
- $\mathcal{X} = \{x_1, \ldots, x_n\}$ is the set of chance nodes represented by circles.
  For any node $n_i \in \mathcal{N}$, $Succ(n_i) \subseteq \mathcal{N}$ denotes the set of its children. Moreover, for any $d_i \in \mathcal{D}, Succ(d_i) \subseteq \mathcal{X}$: $Succ(d_i)$ corresponds to the set of actions that can be decided when d_i is observed. For any $x_i \in \mathcal{X}, Succ(x_i) \subseteq \mathcal{LN} \cup \mathcal{D}$: $Succ(x_i)$ is indeed the set of outcomes of the action x_i - either a leaf node is observed, or a decision node is reached (and then a new action should be executed).

In the present paper, we are interest in the simple case of total ignorance, where information at chance nodes is just a list of potential outcomes without probability distribution.

Solving a decision tree amounts to building a *strategy* δ that selects an action (i.e. a chance node) $\delta(d_i) \in Succ(d_i)$ for each reachable decision node $d_i \in \mathcal{D}$ that associates a chance node $\delta(d_i) \in Succ(d_i)$ to each decision node d_i: $\delta(d_i)$ is the action to be executed when a decision node d_i is reached.

Let Δ be the set of strategies that can be built from the decision tree. Any strategy in Δ can be viewed as a connected subtree of the decision tree where there is exactly one decision arc left at each decision node - skipping the decision nodes, we get a chance tree or, using von Neuwman and Morgernsterm's terminology, a coumpound lottery.

3 Resolute Choice

To evaluate/compare strategies with the Hurwicz criterion, we shall first follow a resolute choice approach. The idea is that any compound lottery is equivalent to a simple one, using a principe of lottery reduction. In our context of decision under total ignorance, no probability distribution is available and reducing a lottery $\mathcal{T}_\delta$ of a strategy δ comes down to computing the multiset of possibly reached states E_δ. We shall now simply compare the strategies by computing $H(\delta) = \alpha \times \min_{s \in E_\delta} u(s) + (1-\alpha) \times \max_{s \in E_\delta} u(s)$ for each of them. This decision-maker behavior is called Resolute Choice and consists in making a strategic decision now and keeping the same strategy over time.

In the literature, the Hurwicz criterion has been generalized to decision trees pervaded with imprecise probabilities by applying it to the reduced lottery [8]. It follows that this criterion will violate Consequentialism in the probabilistic case, and this is also the case in our simple non probabilistic framework. For instance, consider the following decision tree under incomplete information:

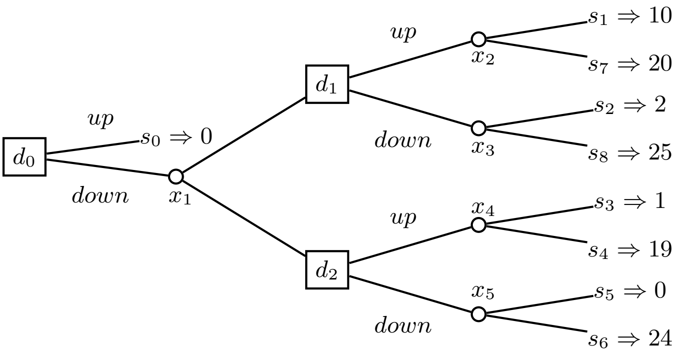


Fig. 1. A Decision Tree

There are 5 strategies: $(d_0 = up)$, $(d_0 = down,\ d_1 = up,\ d_2 = down)$, $(d_0 = down,\ d_1 = down,\ d_2 = up)$, $(d_0 = down,\ d_1 = up,\ d_2 = up)$, $(d_0 = down,\ d_1 = down,\ d_2 = down))$

The problem is to find the best strategy for a Hurwicz decision-maker with degree of optimism α, if there is no other information at x_1, x_2, x_3, x_4, x_5.

The Resolute Choice approach consists in noticing that each strategy yields a different set of possible rewards and computing the best strategy using the prescribed decision criterion.

- Reachable states: $E_{(d_0=down,d_1=up,d_2=up)} = \{s_1, s_7, s_3, s_4\}$: $H(d_0 = down,\ d_1 = up, d_2 = up) = \alpha + 20(1-\alpha)$
- Reachable states: $E_{(d_0=down,d_1=up,d_2=down)} = \{s_1, s_7, s_5, s_6\}$: $H(d_0 = down,\ d_1 = up, d_2 = down) = 24(1-\alpha)$
- Reachable states: $E_{(d_0=down,d_1=down,d_2=up)} = \{s_2, s_8, s_3, s_4\}$: $H(d_0 = down,\ d_1 = down, d_2 = up) = \alpha + 25(1-\alpha)$
- Reachable states: $E_{(d_0=down,d_1=down,d_2=down)} = \{s_2, s_8, s_5, s_6\}$: $H(d_0 = down,\ d_1 = down, down) = 25(1-\alpha)$
- Reachable states: $E_{d_0=up} = \{s_0\}$: $H(d_0 = up) = 0$

So the optimal strategy ex ante consists in deciding for "down" at node d_0, for "down" at node d_1 and "up" at node d_2 what ever $\alpha \in [0,1]$.

However assume Consequentialism. Suppose the decision-maker reaches decision node d_1, because she/he choses "down" at d_0: Now the sets of remaining possible rewards and the corresponding evaluations are for each remaining strategy (the boldface decision is a past one):

- $E_{(\mathbf{d_0}=\mathbf{down},d_1=up)} = \{s_1, s_7\}$: $H(\mathbf{d_0} = \mathbf{down}, up) = 10\alpha + 20(1-\alpha)$
- $E_{(\mathbf{d_0}=\mathbf{down},d_1=down)} = \{s_2, s_8\}$: $H(\mathbf{d_0} = \mathbf{down}, down) = 2\alpha + 25(1-\alpha)$

If the decision-maker is pessimistic enough ($\alpha > \frac{5}{13}$) the best decision d_1 is "up" (and not "down" as found using Resolute Choice). Suppose now that the decision-maker reaches decision node d_2 due to the outcome of chance node x_1:

- $E_{(\mathbf{d_0=up},d_2=up)} = \{s_3, s_4\}$: $H((\mathbf{d_0} = \mathbf{up}, d_2 = up)) = \alpha + 19(1 - \alpha)$
- $E_{(\mathbf{d_0=up},d_2=down)} = \{s_5, s_6\}$: $H(\mathbf{d_0} = \mathbf{up}, d_2 = down)) = 24(1 - \alpha)$

If the decision-maker is enough optimistic $\alpha < \frac{5}{6}$ the best decision d_2 is "down" and not "up" as found using Resolute Choice. In this example is easy to see that for a decision-maker which is not strictly optimistic nor pessimistic (more precisely if $\alpha \in]\frac{5}{13}, \frac{5}{6}[$), both decisions proposed by the Resolute Choice approach are in conflict with the decision-maker's preference at the moments he has to decide.

4 An Alternative to Resolute Choice under Pure Uncertainty

Jaffray [7] starts from the psychological implausibility of the Resolute Choice: an optimal strategy chosen at time t can become unacceptable in the future. This is because in some sense, the decision-maker now is not the same as the decision-maker in the future. Jaffray speaks of different *egos*. He assigns a different ego to each decision node and tries to build a strategy now that is not dominated for the future egos. The question of Jaffray is: how can egos collaborate? Contrary to Resolute Choice where the present ego enforces his preferences to the future ones, we present two alternatives to the Resolute Choice. The first is the direct application of the Veto-process proposed by Jaffray [7]. The second approach is based on the idea that the satisfaction of the present depends on the one of his future egos.

4.1 A Veto-process under Pure Uncertainty

Let Δ_N be a set of possible strategies from node N. The application of the algorithm proposed by Jaffray to our case of pure uncertainty comes down to letting each ego N select those of its possible strategies $\delta \in \Delta_N$ that are optimal according to the Hurwicz criterion applied on the reduction of δ, i.e. on the the leaves of δ. So the algorithm consists in selecting the best substrategy (strict Veto-process), from the last decision nodes to the root decision node by applying lottery reduction and Hurwicz criterion.

Back to our example: suppose $\alpha \in]\frac{5}{13}, \frac{5}{6}[$; the ego for decision d_1 will choose "up" and the ego of decision d_2 will choose "down". So the ego of decision d_0 will have to decide between strategies:

- $\delta_1 = (d_0 = down, d_1 = up, d_2 = down)$: $E_{\delta_1} = \{s_1, s_7, s_5, s_6\}$ and $H(\delta_1) = 24(1 - \alpha)$
- $\delta_2 = (d_0 = up)$: $E_{d_0=up} = \{s_0\}$ and $H(d_0 = up) = 0$

Then the best strategy is δ_1

Algorithm 1. Veto-process under pure uncertainty

```
Input: decision tree $\mathcal{T}$ of depth $p > 1$, optimism coefficient $\alpha$
Output: A strategy $\delta$
foreach node $N$ from the depth $p-1$ to $0$ in $\mathcal{T}$ do
    if $N \in \mathcal{X}$ then
        $\Delta_N \leftarrow \cup_{N' \in Succ(N)} \{\{(N, N')\} \cup \delta : \delta \in \Delta_{N'}\}$
    if $N \in \mathcal{D}$ then
        $\Delta_N \leftarrow \bigcup_{N' \in Succ(N)} \{\Delta_{N'}\}$
    foreach $\delta \in \Delta_N$ do
        $H_\delta \leftarrow \alpha \times \min_{ln_i \in \delta \cap \mathcal{LN}} u(s_i) + (1-\alpha) \times \max_{ln_i \in \delta \cap \mathcal{LN}} u(s_i)$
    $V_{\max} \leftarrow \max_{\delta \in \Delta_N} H_\delta$
    $\Delta_N \leftarrow \{argmax_{\delta \in \Delta_N} H_\delta\}$;
```

In this approach, the Veto of future egos enforces Consequentialism. Since each ego is responsible from the choice of its strategy, the algorithm ensures that a future ego will not deviate from the chosen strategy.

In the Veto-process, the egos are like independent players, so that each player tries to optimize its own criterion ; one can even imagine that each ego N works with its own degree of optimism α_N over the rewards it finally gets, letting $H_\delta \leftarrow \alpha_N \times \min_{ln_i \in \delta \cap \mathcal{LN}} u(s_i) + (1 - \alpha_N) \times \max_{ln_i \in \delta \cap \mathcal{LN}} u(s_i)$ in the algorithm.

But in sequential decision under uncertainty, all players are dependent since they participate of the same decision-maker. Moreover, only one player gets the final reward. So, what is the relevance of assuming independent egos? In the next section we propose a alternative process where the egos are dependent.

4.2 Ego-dependent Process under Pure Uncertainty

Another possibility could be to consider that the preference degrees of one ego is a function of the satisfaction degree of its future egos - and eventually of the egos that receive the final rewards, considering that like decision nodes, leaves are more or less satisfied egos. Moreover the current decision must put the future egos in the best position to be satisfied, until the last ego (i.e., the ego of a leaf) obtains the final reward.

So in this section, we propose a new criterion that combines Hurwicz's idea and the philosophy of dependent satisfaction of the egos. For the ego of the decision node d_i, the worth of decision N (a given posterior chance node) at this node is recursively obtained as:

$$H(N) = \alpha \times \min_{u \in \mathcal{H}(\mathcal{N})} u + (1 - \alpha) \times \max_{u \in \mathcal{H}(\mathcal{N})} u. \qquad (1)$$

where $\mathcal{H}(\mathcal{N}) = \{H(N') : \forall N' \in Succ(N)\}$.

In other words, this approach comes down to recursively replacing any simple lottery by a certainty equivalent $H(N)$ that would provide utility $H(N)$ for sure.

In a sense, we can say that we use a kind of lottery reduction that depends on the parameters of the criterion used (Hurwicz), and more particulary on α. To reduce a compound lottery, we replace all final simple lotteries by their Hurwicz values, and carry on recursively.

Algorithm 2. Ego-dependent process under pure uncertainty

```
Input: decision tree $\mathcal{T}$ of depth $p > 1$, optimism coefficient $\alpha$
Output: A strategy $\delta^*$
foreach node $N$ from the depth $p$ to 0 in $\mathcal{T}$ do
    if $N \in \mathcal{LN}$ then
        $V_N \leftarrow \{u(N)\}$
    if $N \in \mathcal{X}$ then
        $V_N \leftarrow \cup_{N' \in Succ(N)} V_{\Delta_{N'}}$
    if $N \in \mathcal{D}$ then
        foreach $N' \in Succ(N)$ do
            $H(N') \leftarrow \alpha \times \min_{v \in V_{N'}} v + (1-\alpha) \times \max_{v \in V_{N'}} v$
        $V_N \leftarrow \{\max_{N' \in Succ(N)} H(V_{N'})\}$
        $\delta^* \leftarrow \delta^* \cup (N, argmax_{N' \in Succ(N)} H(V_{N'}))$
```

Under this approach, it follows that the optimal strategy δ^* under this criterion is the strategy which maximizes $H^*_N, \forall N \in Succ(d_0)$. It yields an optimal worth H^*, which is the maximal value of H_N over the substrategies starting by chance nodes in $Succ(d_0)$. For any chance node N in $Succ(d_0)$, H_N is the utility value such that all the future decisions are optimal for all the future egos; so this model has several nice properties:

- Any substrategy of an optimal strategy is optimal with respect to the corresponding decision subtree.
- The optimal strategy can be computed by means of dynamic programming from the leaves to the root of the decision tree.

This approach is appealing because it verifies two properties (*Dynamic Consistency* and *Consequentialism*) and fails the other one (*Tree Reduction*). Indeed, we do not consider that a decision tree is necessarily equivalent to a set of reduced lotteries associated to strategies, in the face of total uncertainty: the structure of the decision tree must influence the choice of strategies. Our Algorithm 2 is thus based on dynamic programming, where V_N, and $H(V_N)$ are respectively sets of evaluations and the worth of node N.

Back to our example with $\alpha = \frac{6}{13}$. The egos of leaves get the utility value associated to their states, so no choice is needed. The ego of decision node d_1 will choose "up" ($H_{d_1=up} = \frac{6}{13} \times 10 + (1 - \frac{6}{13}) \times 20 \sim 15.38$ and $H_{d_1=down} \sim 14.38$), the ego of decision node d_2 will choose "down" ($H_{d_2=up} \sim 10.69$ and $H_{d_2=down} \sim 12.92$). The ego of decision node d_0 only has to compare the decision "up" with the decision "down" when $d_1 = up$, and $d_2 = down$ using Hurwicz criterion on aggregated values $\{15.38, 12.92\}$; so the ego of d_0 choice "down".

In the example, the two methods give the same optimal strategy, but they would not if $H_{d_0=up} = 14$ since then the would be optimal using the Veto-Process (it actually considers only the ego at d_2).

5 Conclusion

In this paper we investigate the problem of sequential decision under pure uncertainty. We argue that the resolute Choice is not always psychologically acceptable. Then we study Jaffray's Veto-process in the context of pure uncertainty. We point out a weakness since it considers that preferences of each ego are independent while there is only one decision-maker over the whole sequential decision process. Hence a new rational approach is proposed, under the assumption that egos try to help the next egos in the sequence. An approach which satisfies this rationality requirement is presented along with an algorithm. Our finding suggest that defining the optimal criterion on the basis of a single reduced lottery induced by the whole decision tree is not satisfactory and looks like a debatable fiction that neglects the structure of decision trees. This is in line with the causal approach to probability revived by Shafer [11] who points out that probability trees were considered more expressive than probability distributions in early times of probability theory. In fact, the thesis of this paper is that computing the decision criterion on the reduced lotteries induced by a decision tree is generally not faithful to what is expected from a best strategy in the face of total uncertainty. Indeed, the optimal worth computed by non-expected utility criteria corresponds to no actual reward, it just reflects the decision maker attitude in front of uncertainty. It contrasts with the classical case, where the optimal expected utility of strategies accounts for the actual satisfaction of the decision-maker after playing the optimal strategy a sufficient number of times, if nature acts according to the prescribed probabilities.

References

1. McClennen, E.F.: Rationality and Dynamic choice: Foundational Explorations. Cambridge University Press, Cambridge (1990)
2. Gilboa, I.: A Combination of Expected Utility and Maxmin Decision Criteria. Journal of Mathematical Psychology 32, 405–420 (1988)
3. Grant, S., Kajii, A., Polak, B.: Decomposable Choice under Uncertainty. Journal of Economic Theory 92(2), 169–197 (2000)
4. Hammond, P.: Consequentialist Foundations of Expected Utility. Theory and Decision 25, 25–78 (1988)
5. Hurwicz, L.: Optimality Criteria for Decision Making under Ignorance. Cowles Commission Papers 370 (1951)
6. Jaffray, J.-Y.: Linear Utility Theory for Belief Functions. Operational Research Letters 82, 107–112 (1989)
7. Jaffray, J.-Y.: Rational Decision Making With Imprecise Probabilities. In: Proc. Int. Sympos. on Imprecise Probabilities (ISIPTA 1999), pp. 183–188 (1999)

8. Jeantet, G.: Algorithmes pour la décision séquentielle dans l'incertain: optimisation de l'utilité espérée dépendant du rang et du critère de Hurwicz, Ph. D. Thesis, Université Paris VI (2010)
9. Raiffa, H.: Decision Analysis: Introductory Lectures on Choices Under Uncertainty. Addison-Wesley, Reading (1968)
10. Savage, L.J.: The Foundations of Statistics. Dover Publications, New York (1972)
11. Shafer, G.: The Art of Causal Conjecture. MIT Press (1996)

Comparison of Fuzzy and Crisp Random Variables by Monte Carlo Simulations

Olgierd Hryniewicz

Systems Research Institute,
Newelska 6, 01-447 Warszawa, Poland
hryniewi@ibspan.waw.pl
http://www.ibspan.waw.pl/~hryniewi

Abstract. Fuzzy random variables are used when randomness is merged with imprecision described by fuzzy sets. When we need to use computer simulations for the comparison of a classical probabilistic approach with that based on fuzzy random variables we need to establish the method for the generation of crisp random variables compatible with existing fuzzy data. In the paper we consider this problem, and propose some practical solutions.

Keywords: random variable, comparison of fuzzy and crisp random variables, Monte Carlo simulation.

1 Introduction

Random phenomena are usually modeled by classical probabilistic models. These models are definitely appropriate when sample observations of random variables are precisely reported, even if their actual values are not precisely known or may substantially vary. However, real statistical data may be defined in imprecise way. Firstly, the observed data may be reported using imprecise linguistic terms like "about one hundred" etc. Moreover, in reality there is often no reason to assume that the unknown values of observations are governed by the same probability distribution. In contrast to the case of precisely known observations, there is no method for the statistical verification of this important hypothesis.

To overcome the problems with the analysis of imprecisely reported statistical data two general approaches are used. First approach, still mainly used, is entirely probabilistic. The supporters of this approach propose to use complex, often multi-level, probabilistic models with many assumptions that are hardly verifiable in practice. The second approach is based on the notion of fuzzy random variables. In this approach imprecise observations are described by fuzzy sets such as e.g. fuzzy numbers. Fuzzy random variables have been introduced in order to merge this imprecision with pure randomness.

When we are dealing with complex problems whose formal description involves random variables it is usually not possible to solve these problems in purely analytical way. Therefore, in such cases we use computer simulation methods, known

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_2

as Monte Carlo methods. When fuzziness is additionally involved in the description of complex problems their solution requires the usage of simulation of fuzzy random variables. The paper by Colubi et al. [1] serves as a very good example how simulation techniques may be used for the analysis of the properties of fuzzy random variables. There also exist numerous papers whose authors propose different methods for simulating fuzzy random variables for the solution of practically oriented problems. However, only few of them provide more general information about the methodology of simulation. The most general formal model that can be used for the simulation of fuzzy random variable has been proposed by Gonzalez-Rodriguez et al. [3]. The methodology presented in this paper is based on the general definition of the fuzzy random variable proposed by Puri and Ralescu [7], and the concept of the simulation of random elements in the separable Hilbert space.

When fuzzy random variables are used for modeling imprecisely described random phenomena an important question often arises about the advantage of this methodology over the classical one. The adherents of purely probabilistic approach claim that it is always possible to describe imprecision using classical probabilistic methods. In this paper we claim that in general they are right if we define a fuzzy random variable according to the definition firstly introduced by Kwakernaak [6]. However, the purely probabilistic model of the fuzzy random variable may be extremely complicated. Fuzzy methodology, in our opinion, provides tools for good approximations. It is interesting, however, to compare these approximations with the results provided by restricted (simplified) purely probabilistic models. It seems hardly possible to do such comparisons analytically, but we could do it using Monte Carlo simulations. In order to do so we need methods for the simulation of crisp random variables whose observed values are *compatible* with existing imprecise information. The proposal of a useful method for doing this is the main goal of this paper.

The remaining part of the paper is organized as follows. In Section 2 we discuss some important problems related to the simulation of fuzzy random variables. In Section 3 we present main original results of this paper. We use the concept of the possibility distribution, understood according to the interpretation of Dubois and Prade [2], for the construction of a random mechanism that generates crisp random variables compatible with respective fuzzy ones. In the fourth section we illustrate our results with some examples of simulation experiments. The paper is concluded in the last section of the paper.

2 Monte Carlo Generation of Fuzzy Random Variables

The notion of a fuzzy random variable has been defined in several papers, starting from early works of Zadeh on the fuzzy probability. The first generally accepted definition was introduced in the paper by Kwakernaak [6]. Statistical methods based on Kwakernaak's proposal have been developed in the works of Kruse (see [5]), so nowadays this approach is often coined as Kwakernaak-Kruse approach. The definition, we present below, is taken from [4], and is consistent with the results of Kruse and Kwakernaak.

Suppose that a random experiment is described as usual by a probability space $(\Omega, \mathcal{A}, \mathcal{P})$, where Ω is a set of all possible outcomes of the experiment, $\mathcal{A}$ is a $\sigma-$ algebra of subsets of Ω (the set of all possible events) and P is a probability measure

Definition 1. *A mapping $X : \Omega \to \mathcal{FN}$ is called a fuzzy random variable if it satisfies the following properties:*

(a) $\{X_\alpha(\omega) : \alpha \in [0,1]\}$ is a set (α-cut) representation of $X(\omega)$ for all $\omega \in \Omega$,
(b) for each $\alpha \in [0,1]$ both $X_\alpha^L = X_\alpha^L(\omega) = \inf X_\alpha(\omega)$ and $X_\alpha^U = X_\alpha^U(\omega) = = \sup X_\alpha(\omega)$, are usual real-valued random variables on $(\Omega, \mathcal{A}, \mathcal{P})$.

According to Kruse [5] a fuzzy random variable X may be considered as an *imprecise perception* of an unknown usual random variable $V : \Omega \to \mathcal{R}$, called an *original* of X. There exists a more general definition proposed by Puri and Ralescu [7], but in this paper we restrict our attention to the case of the fuzzy random variable defined according to the Kwakernaak-Kruse approach.

Let us look at the definition of the fuzzy random variable from a point of view of computer simulations. It seems to be quite obvious that the ordinary random variables X_α^L and X_α^U must be dependent. Moreover, for all α-levels $0 \le \alpha \le 1$, and for all pairs of α-levels $0 \le \alpha' \le \alpha'' \le 1$ their joint probability distribution must fulfill the following requirements that assure the nested structure of the α-level subsets of the fuzzy observations.

$$P\left(x_\alpha^L < X_\alpha^L, X_\alpha^U \le x_\alpha^U\right) : \begin{cases} \ge 0 \ , x_\alpha^U > x_\alpha^L \\ = 0 \ , otherwise \end{cases} \tag{1}$$

$$P\left(x_{\alpha'}^L < X_{\alpha'}^L, X_{\alpha''}^L \le x_{\alpha''}^L\right) : \begin{cases} \ge 0 \ , x_{\alpha''}^L > x_{\alpha'}^L \\ = 0 \ , otherwise \end{cases} \tag{2}$$

$$P\left(x_{\alpha''}^U < X_{\alpha''}^U, X_{\alpha'}^U \le x_{\alpha'}^U\right) : \begin{cases} \ge 0 \ , x_{\alpha'}^U > x_{\alpha''}^U \\ = 0 \ , otherwise \end{cases} \tag{3}$$

Thus, we have the following proposition.

Proposition 1. *Let the fuzzy random variable $\tilde{X}$ be defined on a finite set of α levels $0 \le \alpha^{(1)} < \alpha^{(2)} < \cdots < \alpha^{(m)} \le 1$. Then, $\tilde{X}$ is fully described by a $2m$-dimensional vector $\left(X_{\alpha^{(1)}}^L, \ldots, X_{\alpha^{(m)}}^L, X_{\alpha^{(1)}}^U, \ldots, X_{\alpha^{(m)}}^U\right)$ of ordinary random variables whose joint probability distribution fulfills the conditions (1)-(3).*

When the values of α are not discretized the random vector mentioned in Proposition 1 becomes infinitely dimensional. Hence, any fuzzy random variable defined according to the Kwakernaak-Kruse approach can be represented by a fully probabilistic model described by dependent ordinary random variables whose marginal probability distributions must fulfill conditions(1)-(3). This property of the fuzzy random variables fully justifies the usage of Monte Carlo methods for the generation of fuzzy random samples.

3 Monte Carlo Simulation of Random Variables Compatible with Fuzzy Data

Let us assume that we have a sample of imprecise fuzzy data $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ observed in a random experiment (or simulated as the realizations of fuzzy random variable). If we want to compare fuzzy models and crisp probabilistic models using simulation methods we need to assume a certain probabilistic model in order to generate possible "origins" of the observed fuzzy data. The most frequently used approach consists in the transformation of the membership functions of the fuzzy numbers $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ into respective probability densities. For example, let us suppose that the fuzzy observation $\tilde{x_i}$ is described by a triangular normal fuzzy number (x_{i1}, x_{i2}, x_{i3}) (i.e. such that $\mu(x_{i2}) = 1$). Then, its membership function is easily transformed to the triangular probability density described by the triangle (x_{i1}, x_{i2}, x_{i3}) such that $f(x_{i2}) = 2/(x_{i3} - x_{i1})$.

This simple model has one serious disadvantage. As a matter of fact, it is a purely probabilistic model that fits the imprecise data to one specific probability distribution. We believe that this assumption is debatable. Consider, for example, two random fuzzy variables $\tilde{X}$ and $\tilde{Y}$ whose observations are described by intervals (i.e. by rectangular fuzzy numbers). According to the theory of fuzzy sets the observations of their sum should be also described by intervals. However, the probability distribution of the sum of their crisp "origins" simulated using the aforementioned method is not uniform.

In this paper we propose to interpret the membership functions of the observed fuzzy data as *possibility distributions*. The notion of the possibility distribution was introduced by Zadeh, and has many different interpretations. According to one of them, see [2], the possibility distribution can be understood as an upper envelope for all ordinary discrete probability distributions compatible with our imprecisely described value. Let $\mu_{[a,b]}(x)$ be the membership function of a fuzzy number $\tilde{x}$ with the support $[a, b]$. Consider now a representation of $[a, b]$ with a finite set of m real numbers $a \leq x_1 < x_2 < \ldots < x_m \leq b$. Now, let us define on this set the family $\mathcal{MN}$ of all *discrete* distributions $MN(p_1, p_2, \ldots, p_m)$ such that $p_j \leq \mu_{[a,b]}(x_j), j = 1, \ldots, m$, and $\sum_{j=1}^{m} p_i = 1$. The discrete distribution that belongs to the family $\mathcal{MN}$ we will call *compatible* with the possibility distribution $\mu_{[a,b]}(x)$. The value $x^\star$ randomly generated according to this distribution can be considered as a possible crisp "origin" of the fuzzy observation $\tilde{x}$.

For every fuzzy number $\tilde{x}$ defined on a non-degenerate interval $[a, b]$ there exist uncountably many distributions defined in the aforementioned way. However, for practical reasons we have to restrict the number of considered distributions. First, we should set the fixed number of points m. When $\mu_{[a,b]}(a) > 0$ and $\mu_{[a,b]}(b) > 0$ we set, respectively, $x_1 = a$ and $x_m = b$, and the remaining $m - 2$ points we may generate according to a certain probability distribution defined on (a, b). Otherwise, we generate all m points from this distribution. Note that any non-random generation of these m points (e.g. equidistant) can be considered as a special case of this general model.

Now, let us define some distributions on the set $\{x_1, \ldots, x_m\}$ that belong to the family $\mathcal{MN}$. Let us consider three such distributions: the *left concentrated* (LC), the *right concentrated* (RC), and the *random Dirichlet* (RD). The LC distribution is the distribution $MN(p_1, \ldots, p_k)$, where $p_j = \mu_{[a,b]}(x_j), j = 1, \ldots, k-1$ and $p_k = 1 - \sum_{j=1}^{k-1} p_j, 1 \geq k \leq m$. The respective RC distribution is the distribution $MN(p_l, \ldots, p_m)$, where $p_j = \mu_{[a,b]}(x_j), j = l+1, \ldots, m$ and $p_l = 1 - \sum_{j=l+1}^{m} p_j, 1 \geq l \leq m$. The interpretation of these distributions is simple when we are interested in the inference about the location parameter of the considered probability distribution or when the parameter of interest could be transformed to a location parameter by the appropriate transformation of the underlying random variable. For purely interval data the whole probability mass of the LC distribution is concentrated at the left limiting value of the considered interval. Similarly,for the RC distribution the whole probability mass is concentrated at the right limiting value of the considered interval.

Let us consider the LC distribution compatible with the triangular possibility distribution defined by the triangular fuzzy number (A, B, C). Let $s = |A, C|$ and $s_L = |A, B|$ be, respectively, the support and the left spread of this possibility distribution. We can now formulate the following proposition.

Proposition 2. *The expected value of the LC distribution compatible with triangular possibility distribution (A, B, C), and defined on the m evenly distributed points on the interval (A, C) is equal to A when m tends to infinity.*

Proof. Let X be a random variable defined on points $x_i = \frac{2i-1}{2}\frac{s}{s_L} + A$ evenly distributed on (A, C), and $p_i = \frac{2i-1}{2m}\frac{s}{s_L}$ be the corresponding probabilities of the LC distribution, such that $\sum_{i=1}^{k} p_i \leq 1 < \sum_{i=1}^{k+1} p_i$. One can prove that

$$Z_k = \sum_{i=1}^{k} \frac{2i-1}{2m}\frac{s}{s_L} = \frac{s}{2ms_L}k^2 \tag{4}$$

Hence, we have $p_{k+1} = 1 - Z_k$. If probabilities $p_1, p_2, \ldots, p_k, p_{k+1}$ describe the probability distribution defined on the set $\{x_1, x_2, \ldots, x_k, x_{k+1}\}$ the condition $Z_k + p_{k+1} = 1$ must be fulfilled. Note that $p_{k+1} \leq \frac{2k+1}{2m}\frac{s}{s_L}$, and for m tending to infinity this probability tends to zero. Thus, we have the following equation that defines the relationship between k and m

$$\frac{s}{2ms_L}k^2 = 1 \tag{5}$$

Now, let us calculate the expected value of X from the following formula

$$\begin{aligned} E(X) &= \textstyle\sum_{i=1}^{k} p_i x_i = \sum_{i=1}^{k} (\frac{2i-1}{2}\frac{s}{m} + A)(\frac{2i-1}{2m}\frac{s}{s_L}) \\ &= \frac{s^2}{s_L}\frac{1}{4m^2}\frac{1}{3}k(4k^2-1) + \frac{A}{2m}\frac{s}{s_L}k^2. \end{aligned} \tag{6}$$

From (5) we have $k = \sqrt{2ms_L/s}$, and hence

$$E(X) = \frac{s^{3/2}}{s_L^{1/2}}\frac{\sqrt{2}}{12}(\frac{8s_L}{s}m^{-1/2} - m^{-3/2}) + A. \tag{7}$$

Thus, from (7) we see that for $m \to \infty$ we have $E(X) \to A$, and this ends the proof. In the similar way we can prove a symmetric proposition concerning the RC distribution

Proposition 3. *The expected value of the RC distribution compatible with triangular possibility distribution* (A, B, C)*, and defined on the* m *evenly distributed points on the interval* (A, C) *is equal to* C *when* m *tends to infinity.*

In the general case, however, the values of the parameters $p_1, \ldots, p_m$ may be also chosen in a random way. The Dirichlet distribution is defined by the following density function

$$f(p_1, \ldots, p_m; \beta_1, \ldots, \beta_m) = \begin{cases} \frac{1}{B_m} \prod_{j=1}^{m} s_j^{\beta_j} & , (s_1, \ldots, p_m) \in S_m \\ \quad 0 & , \quad otherwise \end{cases} , \tag{8}$$

where S_m is the closed $m-1$-dimensional simplex and B_m is the normalizing constant.

The Dirichlet distribution defined by (8) is very flexible, and allows to simulate very different "shapes" of the probability distribution compatible with a fuzzy number $\tilde{x}$ which can be used for the generation of the "origin" value representative for this fuzzy number. In the simulation algorithm the values of the parameters $\beta_1, \ldots, \beta_m$ can be chosen randomly, for example from a predefined interval $(\beta_{min}, \beta_{max})$. It gives an additional level of flexibility in the generation of probabilities $p_1, \ldots, p_m$. Then, the values of probabilities $p_1, \ldots, p_m$ can be generated from the Dirichlet distribution (8). Finally, the $MN(p_1, \ldots, p_m)$ distribution can be used for the generation of the "origin" of the observed value of the fuzzy random variable from among the set of (predefined or randomly generated) values $\{x_1, \ldots, x_m\}$. We have called this distribution the random Dirichlet (RD).

4 Properties of Probability Distributions Representing Fuzzy Random Variables – Results of Experiments

From the discussion presented in Section 2 we know that a fuzzy random observation $\tilde{z}$ can be represented as the sum of the unobserved crisp random "origin" y and a fuzzy number $\tilde{x}$ that represents our lack of knowledge about the "origin". In Monte Carlo experiments we can simulate "origins" from a given probability distribution. Then, we can use a certain predefined random mechanism for the generation of the membership function $\mu(x)$ of $\tilde{x}$. Thus, the simulated fuzzy observation is a fuzzy number $\tilde{z} = y + \tilde{x}$.

When we need to compare the approach based on random fuzzy numbers with a classical approach based on crisp random numbers we should simulate crisp random numbers that are compatible with our fuzzy observations. In this section we present the results of simulation experiments that have been performed in order to investigate the differences between different methods of the simulation of random variables compatible with given fuzzy observations. In this paper we describe only the results of experiments in which we assumed that the observed fuzzy numbers are described by triangular membership functions symmetric around zero with the randomly generated left and right spreads L and R. Moreover, we assumed that both are described by the same probability distribution characterized by its expected value w and a coefficient of variation v.

In our experiments we have considered four types of probability distributions compatible with the given fuzzy observation: triangular, left(right) concentrated, and random Dirichlet. Because the spreads L and R have been simulated from the same distribution the average behavior of the LC and RC distributions was the same. Therefore, we present here only the results for the RC distribution. In the experiments we generated samples of n random fuzzy variables $\tilde{X}$, Then, for each generated sample item we generated its "origin" from its compatible probability distribution. In the next step we calculated the sample average of the generated "origins". The procedure has been repeated 100 000 times in order to evaluate the properties of the simulated distributions of sample averages, such as the expected value and the standard deviation.

In the first group of simulation experiments we investigated the dependence of the expected value of the RC (LC) distribution on the number of discretization points m when the triangular membership functions were generated from different probability distributions. The convergence to the limiting values defined by Proposition 3 (or 2) was rather slow. For example, when the spreads were generated from the uniform distribution defined on the interval $(0, 4)$ the expected value for the RC distribution for $m = 500$ was equal to $1, 88$. Note however, that according to Proposition 3 for $m \to \infty$ this expected value should be equal to 2, i.e. to the expected length of the right spread. The results obtained in similar experiments have shown that for a realistic discretization of the possibility distribution described by a fuzzy number the observed average values of the left and right concentrated distributions are not so far from their theoretical, but rather improbable, values.

The important question may arise about the difference between the random variables generated from the triangular (Tr) distribution and the random variables generated with the usage of the proposed random Dirichlet (RD) distribution. Because of the way the triangular membership functions are generated (maximum at zero, the same distribution of the both, left and right, spreads) the expected value of the sample average for this distribution must be equal to zero. However, in the case of the random Dirichlet distribution the similar behavior of the sample average is somewhat unexpected. Only in the case of small values of m the estimated average is slightly different than zero (e.g. for $m = 10$ it is equal to -0.046 while σ is equal to 0.265). The situation is different when we consider the standard deviation of $\bar{x}$.

In all considered cases the variability of sample means generated from the random Dirichlet distribution was greater than the similar variability in the case of the triangular distribution. This difference becomes significant for moderate and large values of m (e.g. larger than 200). This means that the random Dirichlet distribution is less "informative", and represents fuzziness in a better way.

Moreover, for the large values of m the standard deviation of $\bar{x}$ practically does not depend upon the value of this parameter. Therefore, the maximal variability of the generated crisp observations that are compatible with the given fuzzy number can be obtained even for a moderate number of discretization points. This property tells us that in real experiments the value of m need not be too large, and thus, simulation experiments need not be time consuming. One should also remember that in practice the variability of the distribution compatible with fuzzy observations is equal to the sum of the variability of an unobserved "origin" and the variability of the distribution representing observed fuzziness. When the former is much larger than the latter the difference between the triangular and the random Dirichlet distributions may be neglected.

5 Conclusions

The widely used methods of the generation of fuzzy random variables are fully compatible with the Kwakernaak-Kruse definition of the fuzzy random variable. The concept of the probability distribution compatible with a fuzzy observation introduced in this paper provides a simple methodology for the comparison of classical (non-fuzzy) and fuzzy approaches for dealing with imprecise data. In the classical approach the lack of knowledge is modeled by a predefined and difficult to identify probability distribution. When the fuzzy approach is used this lack of knowledge may be modeled by several probability distributions that are compatible, in the sense introduced in the paper, with imprecise observations. Therefore, this approach provides more flexibility in the description of imprecisely observed random phenomena.

Acknowledgments. The author thanks Dr. Piotr Nowak for his helpful comments regarding the interpretation of fuzzy random variables.

References

1. Colubi, A., Fernandez-Garcia, C., Gil, M.A.: An Empirical Approach to Statistical/Probabilistic Studies With Fuzzy Experimental Data. IEEE Transactions on Fuzzy Systems 10, 384–390 (2002)
2. Dubois, D., Prade, H.: Possibility Theory. Plenum Press, New York (1988)
3. Gonzalez-Rodriguez, G., Colubi, A., Trutschnig, W.: Simulation of fuzzy random variables. Information Sciences 179, 642–653 (2008)
4. Grzegorzewski, P., Hryniewicz, O.: Computing with words and life data. Int. Journ. of Appl. Math. & Comp. Science 12, 337–345 (2002)
5. Kruse, R., Meyer, K.D.: Statistics with Vague Data. Riedel, Dordrecht (1987)
6. Kwakernaak, H.: Fuzzy random variables, part I: definitions and theorems. Information Sciences 15, 1–15 (1978)
7. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. Journ. of Math. Anal. and Appl. 114, 409–422 (1986)

Stochastic Orders for Fuzzy Random Variables

Ignacio Montes, Enrique Miranda, and Susana Montes

Dept. of Statistics and O.R., University of Oviedo, Spain
{imontes,mirandaenrique,montes}@uniovi.es

Abstract. The comparison of random variables can be made by means of stochastic orders such as expected utility or statistical preference. One possible model when the random variables are imprecisely observed is to consider fuzzy random variables, so that the images become fuzzy sets. This paper proposes two comparison methods for fuzzy random variables: one based on fuzzy rankings and another one that uses the extensions of stochastic orders to an imprecise framework. The particular case where the images of the fuzzy random variables are triangular fuzzy numbers is investigated.We illustrate our results by means of a decision making problem.

Keywords: Fuzzy random variables, stochastic orders, expected utility, statistical preference, possibility measures.

1 Introduction

A decision making problem under uncertainty requires the choice between several alternatives that are usually modeled by means of random variables; the choice between them is made by means of stochastic orders [11]. When we have imprecise information about the consequences of the different alternatives, we need to consider a more general model, such as sets of random variables, random sets, or, as we do in this paper, fuzzy random variables [6], where the images are fuzzy sets instead of real numbers. In order to extend stochastic orderings to this case, we follow in this paper two different avenues. On the one hand, based on the idea behind statistical preference, we compare fuzzy random variables by means of a choice model over their images, using fuzzy rankings, where by *fuzzy ranking* we refer to a method for the comparison of fuzzy sets. On the other hand, and similarly to expected utility, we can also compare fuzzy random variables in terms of their expectations. Since the expectation of a fuzzy random variable can be modeled by a possibility measure, we shall use the methods established in [9,10] for the comparison of imprecise probability models.

The paper is organized as follows: Section 2 introduces the main notions about fuzzy random variables and stochastic orders defined under imprecision. Then we discuss the two approaches mentioned above for the comparison of fuzzy random variables, and in Section 4 we investigate the particular case where the images of the fuzzy random variables are triangular fuzzy numbers. Finally, Section 5 illustrates our methods in a decision making problem. The paper concludes with

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_3

some additional remarks and a discussion of other approaches to this problem. Proofs are omitted because of space limitations.

2 Preliminary Notions

2.1 Fuzzy Random Variables

Fuzzy random variables were introduced simultaneously by Kruse and Meyer [6] and Puri and Ralescu [12]. In this paper, we follow the epistemic approach considered in [6]. Let $\mathcal{F}(\mathbb{R})$ denote the set of all fuzzy sets on $\mathbb{R}$.

Definition 1 ([6]). *A fuzzy random variable is a map $\widetilde{X} : \Omega \to \mathcal{F}(\mathbb{R})$ such that the α-cuts $\widetilde{X}_\alpha$ are strongly measurable multi-valued mappings.*

Following Kruse and Meyer, fuzzy random variables can be used to model the imprecise knowledge about an unknown random variable U_0. For any $\omega \in \Omega, \omega' \in \mathbb{R}$, $\widetilde{X}(\omega)(\omega')$ can be interpreted as the acceptability degree of the proposition "$U_0(\omega) = \omega'$". With a similar reasoning, it is possible to define a fuzzy set on the class of measurable functions from Ω to $\mathbb{R}$, $\mu_{\widetilde{X}}$, that associates the value

$$\mu_{\widetilde{X}}(U) = \inf\{\widetilde{X}(\omega)(U(\omega)) : \omega \in \Omega\}$$

for any measurable function $U : \Omega \to \mathbb{R}$. Then, according to [6] this value can be understood as the acceptability degree of the proposition "$U = U_0$". Using the fuzzy set $\mu_{\widetilde{X}}$ it is possible to define the expectation of a fuzzy random variable as the fuzzy set $E_{\widetilde{X}}$ with membership function:

$$E_{\widetilde{X}}(r) = \sup\{\mu_{\widetilde{X}}(U) : E(U) = r\}. \tag{1}$$

$E_{\widetilde{X}}(r)$ can be interpreted as the acceptability degree of the proposition "$E(U_0) = r$". This membership function can also be seen as a possibility distribution, and as a consequence this expectation can be regarded as a possibility measure.

2.2 Stochastic Orders under Imprecision

Stochastic orders are methods for the comparison of random quantities. Here we shall use *expected utility*, given by $X \succeq_{\text{EU}} Y \Leftrightarrow E(X) \geq E(Y)$, and *statistical preference* [2,3], that is based on a probabilistic relation. A probabilistic relation on a set of alternatives $\mathcal{A}$ is a map defined from $\mathcal{A}^2$ to $[0,1]$ such that $Q(a,b) + Q(b,a) = 1$ for any $(a,b) \in \mathcal{A}^2$, where $Q(a,b)$ measures the strength of the preference of a over b. Statistical preference considers a set of alternatives formed by random variables, and defines a probabilistic relation by $Q(X,Y) = P(X > Y) + \frac{1}{2}P(X = Y)$. Then, X is statistically preferred to Y, denoted by $X \succeq_{\text{SP}} Y$, if $Q(X,Y) \geq \frac{1}{2}$. In what remains we will use a well-known alternative expression for statistical preference: $X \succeq_{\text{SP}} Y$ if and only if $P(X \geq Y) \geq P(Y \geq X)$.

In a context of imprecision, it may be necessary to choose between *sets* of random variables, instead of single ones. This problem was studied in some detail in [9,10], and a number of extensions of a given stochastic order to the imprecise case were considered. In the next definition, $\succeq$ denotes a stochastic order that could be either the expected utility or statistical preference, as we shall use in this paper, or any other stochastic order.

Definition 2 ([10, Def. 5]). *Consider two sets of random variables $\mathcal{X}, \mathcal{Y}$ and a stochastic order $\succeq$. We say that:*

- *$\mathcal{X}$ is $\succeq_1$-preferred to $\mathcal{Y}$ if $U \succeq V$ for any $U \in \mathcal{X}$ and $V \in \mathcal{Y}$.*
- *$\mathcal{X}$ is $\succeq_2$-preferred to $\mathcal{Y}$ if there is $U \in \mathcal{X}$ such that $U \succeq V$ for any $V \in \mathcal{Y}$.*
- *$\mathcal{X}$ is $\succeq_3$-preferred to $\mathcal{Y}$ if for any $V \in \mathcal{Y}$ there is $U \in \mathcal{X}$ such that $U \succeq V$.*
- *$\mathcal{X}$ is $\succeq_4$-preferred to $\mathcal{Y}$ if there are $U \in \mathcal{X}$ and $V \in \mathcal{Y}$ such that $U \succeq V$.*
- *$\mathcal{X}$ is $\succeq_5$-preferred to $\mathcal{Y}$ if there is $V \in \mathcal{Y}$ such that $U \succeq V$ for any $U \in \mathcal{X}$.*
- *$\mathcal{X}$ is $\succeq_6$-preferred to $\mathcal{Y}$ if for any $U \in \mathcal{X}$ there is $V \in \mathcal{Y}$ such that $U \succeq V$.*

When the extended stochastic order is either expected utility or statistical preference, we shall use the notation $\succeq_{\mathrm{EU_i}}$ or $\succeq_{\mathrm{SP_i}}$, respectively.

Some stochastic orders, such as expected utility, compare two random variables by means of their associated probability distributions. For those, the definitions above can be used to compare sets of probability distributions, also called *credal sets.* This allows us to compare imprecise probability models, such as possibility measures. Indeed, the credal set associated with a possibility measure Π is given by:

$$\mathcal{M}(\Pi) = \{P \text{ probability} \mid P \leq \Pi\}.$$

Then, we can compare two possibility measures Π_{X} and Π_{Y} by means of their associated credal sets. Our next result considers the extensions of expected utility, and uses $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU_i}} \Pi_{\mathrm{Y}}$ to denote $\mathcal{M}(\Pi_{\mathrm{X}}) \succeq_{\mathrm{EU_i}} \mathcal{M}(\Pi_{\mathrm{Y}})$ for $i = 1, \ldots, 6$. Recall also that the conjugate function N of a possibility measure Π, given by $N(A) = 1 - \Pi(A^c)$ for every A, is usually named *necessity measure.*

Proposition 1. *For any two possibility measures Π_{X} and Π_{Y}, with conjugate necessity measures N_{X} and N_{Y}, respectively, it holds that:*

- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU_1}} \Pi_{\mathrm{Y}} \Leftrightarrow (C)\displaystyle\int id \mathrm{d}\Pi_{\mathrm{X}} \geq (C)\int id \mathrm{d}N_{\mathrm{Y}};$
- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU_2}} \Pi_{\mathrm{Y}} \Leftrightarrow \Pi_{\mathrm{X}} \succeq_{\mathrm{EU_3}} \Pi_{\mathrm{Y}} \Leftrightarrow (C)\displaystyle\int id \mathrm{d}N_{\mathrm{X}} \geq (C)\int id \mathrm{d}N_{\mathrm{Y}};$
- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU_4}} \Pi_{\mathrm{Y}} \Leftrightarrow (C)\displaystyle\int id \mathrm{d}N_{\mathrm{X}} \geq (C)\int id \mathrm{d}\Pi_{\mathrm{Y}};$
- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU_5}} \Pi_{\mathrm{Y}} \Leftrightarrow \Pi_{\mathrm{X}} \succeq_{\mathrm{EU_6}} \Pi_{\mathrm{Y}} \Leftrightarrow (C)\displaystyle\int id \mathrm{d}\Pi_{\mathrm{X}} \geq (C)\int id \mathrm{d}\Pi_{\mathrm{Y}};$

where $(C)\int f \mathrm{d}\mu$ denotes the Choquet integral of f with respect to the non-additive measure μ, and id denotes the identity function $id(x) = x$.

3 Comparison of Fuzzy Random Variables

As we mentioned in Section 2.2, two possible ways of comparing two random variables X, Y are expected utility and statistical preference, given by:

$$X \succeq_{\mathrm{EU}} Y \Leftrightarrow E(X) \geq E(Y). \tag{2}$$

$$X \succeq_{\mathrm{SP}} Y \Leftrightarrow P(\{\omega : X(\omega) \geq Y(\omega)\}) \geq P(\{\omega : Y(\omega) \geq X(\omega)\}). \tag{3}$$

In this section, we extend these two orders to fuzzy random variables. In the case of expected utility, the comparison of the expectations leads us to the comparison of possibility measures; concerning statistical preference, the comparison of the images of fuzzy random variables motivates the use of fuzzy rankings.

3.1 Comparison by Means of Fuzzy Rankings

Fuzzy rankings are methods for the comparison of quantities modeled by means of fuzzy sets, in that they measure to what extent one fuzzy set is larger than the other. Consider two fuzzy random variables, $\widetilde{X}, \widetilde{Y}$ modeling the imprecise knowledge of respective random variables X, Y. Then for every ω in the initial space $\widetilde{X}(\omega)$ and $\widetilde{Y}(\omega)$ are the fuzzy sets that represent the degree of acceptability of the propositions "$X(\omega) = \omega'$" and "$Y(\omega) = \omega'$", for any $\omega' \in \mathbb{R}$. In order to compare the fuzzy random variables $\widetilde{X}$ and $\widetilde{Y}$, we can compare the fuzzy sets $\widetilde{X}(\omega)$ and $\widetilde{Y}(\omega)$ for every $\omega \in \Omega$. This leads at once to the following definition:

Definition 3. *Let $\widetilde{X}, \widetilde{Y} : \Omega \to \mathcal{F}(\mathbb{R})$ be two fuzzy random variables on a probability space $(\Omega, \mathcal{A}, P)$, and let $\succsim$ be a fuzzy ranking. We say that $\widetilde{X}$ is $\succsim$-*statistically preferred *to $\widetilde{Y}$, and denote it $\widetilde{X} \succsim^{P} \widetilde{Y}$, when*

$$P(\{\omega \in \Omega : \widetilde{X}(\omega) \succsim \widetilde{Y}(\omega)\}) \geq P(\{\omega \in \Omega : \widetilde{Y}(\omega) \succsim \widetilde{X}(\omega)\}).$$

When the fuzzy ranking $\succsim$ is complete, that is, if it allows the comparison of every pair of fuzzy sets, we obtain the following result.

Proposition 2. *Let $\succsim$ be a complete fuzzy ranking, and define:*

$$Q(\widetilde{X}, \widetilde{Y}) = P(\{\omega : \widetilde{X}(\omega) \succ \widetilde{Y}(\omega)\}) + \frac{1}{2} P(\{\omega : \widetilde{X}(\omega) \sim \widetilde{Y}(\omega)\}).$$

Then $Q(\widetilde{X}, \widetilde{Y}) + Q(\widetilde{Y}, \widetilde{X}) = 1 \ \forall \widetilde{X}, \widetilde{Y}$, and $\widetilde{X}$ is $\succsim$-statistically preferred to $\widetilde{Y}$ if and only if $Q(\widetilde{X}, \widetilde{Y}) \geq \frac{1}{2}$. Moreover, if $\succsim$ extends the natural order on $\mathbb{R}$, then $\succsim$-statistical preference is an extension of statistical preference given by Eq. (3).

3.2 Comparison by Means of Stochastic Orders

Another way of comparing fuzzy random variables is by extending appropriately the order associated with expected utility, given by Eq. (2). Consider two fuzzy random variables $\widetilde{X}$ and $\widetilde{Y}$, and let $E_{\widetilde{X}}, E_{\widetilde{Y}}$ be their respective fuzzy expectations, given by Eq. (1). These expectations are fuzzy sets, or, equivalently, possibility measures. It leads to the following definition.

Definition 4. *We say that $\widetilde{X}$ is preferred to $\widetilde{Y}$ with respect to the i-th extension of expected utility, and denote it $\widetilde{X} \succeq_{\mathrm{EU_i}} \widetilde{Y}$, when $E_{\widetilde{X}} \succeq_{\mathrm{EU_i}} E_{\widetilde{Y}}$, where $\succeq_{\mathrm{EU_i}}$ is given in Definition 2.*

This result, together with Proposition 1, reduces the comparison of fuzzy random variables to the comparison of appropriate Choquet integrals. For a thorough discussion of the interpretation behind the different extensions $\succeq_{\mathrm{EU_i}}$, for $i = 1, \ldots, 6$, we refer to [9,10].

4 Particular Case: Triangular Fuzzy Random Variables

In this section we study the particular case where the images of $\widetilde{X}$ and $\widetilde{Y}$ are triangular fuzzy numbers. Recall that $A = (a_1, a_2, a_3)$ is a *triangular fuzzy number* when its membership function is given by:

$$A(\omega) = \begin{cases} \frac{x-a_1}{a_2-a_1} & \text{for } a_1 < x \leq a_2. \\ \frac{a_3-x}{a_3-a_2} & \text{for } a_2 < x \leq a_3. \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

4.1 Fuzzy Rankings on Triangular Fuzzy Random Variables

Fuzzy rankings usually take a simple expression when applied to triangular fuzzy numbers. Here we consider two well-known fuzzy rankings, introduced by Dubois and Prade in [5].

Definition 5 ([5]). *Let A, B be two fuzzy numbers, and define:*

- **Possibility of Dominance:** $PD(A, B) = \sup\limits_{x \geq y}(\min(A(x), B(y)))$.
- **Necessity of Strict Dominance:** $NSD(A, B) = 1 - \sup\limits_{x \leq y}(\min(A(x), B(y)))$.

Then we denote $A \succeq_{\mathrm{PD}} B$ when $PD(A, B) \geq PD(B, A)$ (and similarly for NSD). In case of triangular fuzzy numbers, these definitions can be simplified:

Lemma 1. *Let $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$ be two triangular fuzzy numbers. It holds that $A \succeq_{\mathrm{PD}} B \Leftrightarrow A \succeq_{\mathrm{NSD}} B \Leftrightarrow a_2 \geq b_2$.*

Proof. This is a consequence of Eq. (4) and Definition 5.

Using this result, we can simplify Definition 3 for these fuzzy rankings.

Proposition 3. *Given two triangular fuzzy random variables $\widetilde{X}$ and $\widetilde{Y}$ such that $\widetilde{X}(\omega) = (a_1^{\omega}, a_2^{\omega}, a_3^{\omega})$ and $\widetilde{Y}(\omega) = (b_1^{\omega}, b_2^{\omega}, b_3^{\omega})$ $\forall \omega \in \Omega$,*

$$\widetilde{X} \succsim^{P}_{\mathrm{PD}} \widetilde{Y} \Leftrightarrow \widetilde{X} \succsim^{P}_{\mathrm{NSD}} \widetilde{Y} \Leftrightarrow P(\{\omega \in \Omega : a_2^{\omega} \geq b_2^{\omega}\}) \geq P(\{\omega \in \Omega : b_2^{\omega} \geq a_2^{\omega}\}).$$

Note also that both PD and NSD are complete fuzzy rankings, and then Proposition 2 can be applied.

4.2 Stochastic Orders on Triangular Fuzzy Random Variables

We now turn on the comparison of triangular fuzzy random variables by means of the generalizations of expected utility. We begin by showing a well-known result that easily allows to compute the expectation of a triangular fuzzy number.

Proposition 4 ([4,7]). *Consider a fuzzy random variable $\widetilde{X}$ such that $\widetilde{X}(\omega)$ is a triangular fuzzy number $(a_1^\omega, a_2^\omega, a_3^\omega)$ for any ω. Consider the functions $f_1(\omega) = a_1^\omega$, $f_2(\omega) = a_2^\omega$ and $f_3(\omega) = a_3^\omega$, for any $\omega \in \Omega$. Then, $E_{\widetilde{X}} = (e_1, e_2, e_3)$ is also a triangular fuzzy number, where $e_1 = E(f_1)$, $e_2 = E(f_2)$ and $e_3 = E(f_3)$.*

Next we show that Definition 4 can be simplified in this case. The proof follows by considering the interpretations of Definition 2 in the case of expected utility (see for instance [9, Remark 3]), and the formulas of the 'best' and 'worst' alternatives in the credal set associated with a possibility measure in the particular case of triangular fuzzy numbers.

Proposition 5. *Consider two possibility measures Π_{X} and Π_{Y} whose associated fuzzy sets are the triangular fuzzy numbers (a_1, a_2, a_3) and (b_1, b_2, b_3), respectively. Then:*

- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU}_1} \Pi_{\mathrm{Y}} \Leftrightarrow a_1 + a_2 \geq b_2 + b_3$.
- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU}_2} \Pi_{\mathrm{Y}} \Leftrightarrow \Pi_{\mathrm{X}} \succeq_{\mathrm{EU}_3} \Pi_{\mathrm{Y}} \Leftrightarrow a_2 + a_3 \geq b_2 + b_3$.
- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU}_4} \Pi_{\mathrm{Y}} \Leftrightarrow a_2 + a_3 \geq b_1 + b_2$.
- $\Pi_{\mathrm{X}} \succeq_{\mathrm{EU}_5} \Pi_{\mathrm{Y}} \Leftrightarrow \Pi_{\mathrm{X}} \succeq_{\mathrm{EU}_6} \Pi_{\mathrm{Y}} \Leftrightarrow a_1 + a_2 \geq b_1 + b_2$.

5 Example of Application in Decision Making

This section presents an application of the previous definitions to a decision making problem. We use the setting considered in [8]: a company operating in UK is considering the possibility of expanding to new markets. They consider four alternatives:

A₁: Expand to the French market.
A₂: Expand to the German market.
A₃: Expand to the Italian market.
A₄: Expand to the Spanish market.

The evaluation of the strategies depends on the economic situation for the next year, that may take four different values:

S₁: Bad economic situation.
S₂: Regular economic situation.
S₃: Good economic situation.
S₄: Very good economic situation.

The probabilities for each state are estimated as 0.1, 0.3, 0.3 and 0.3, respectively. Then, we can define the probability space $(\Omega, \mathcal{P}(\Omega), P)$, where $\Omega = \{S_1, S_2, S_3, S_4\}$, and model each alternative as a fuzzy random variable taking the following values, that represent the expected benefits:

	S_1	S_2	S_3	S_4
A_1	$(0.2, 0.3, 0.4)$	$(0.6, 0.7, 0.8)$	$(0.2, 0.3, 0.4)$	$(0.5, 0.6, 0.7)$
A_2	$(0.5, 0.5, 0.5)$	$(0.3, 0.4, 0.5)$	$(0.4, 0.5, 0.7)$	$(0.4, 0.5, 0.6)$
A_3	$(0.1, 0.2, 0.4)$	$(0.6, 0.8, 0.9)$	$(0.8, 0.9, 1)$	$(0.7, 0.8, 0.9)$
A_4	$(0.3, 0.4, 0.5)$	$(0.3, 0.4, 0.6)$	$(0.5, 0.5, 0.5)$	$(0.3, 0.4, 0.5)$

Since these alternatives are triangular fuzzy random variables, we can apply the results from Section 4. First of all, if we compare them pairwisely by means of PD and NSD, Lemma 1 assures that the two fuzzy rankings reduce to the comparison of the modal points of the triangular fuzzy numbers. The resulting preference degrees are summarized in the following table:

	A_1	A_2	A_3	A_4
A_1	·	0.5	0.25	0.5
A_2	0.5	·	0.25	0.75
A_3	0.75	0.75	·	1
A_4	0.5	0.25	0	·

Thus, we conclude that the best alternative is A_3, that is, to invest into the Italian market. If instead we compare these fuzzy random variables by means of the generalized expected utility, we deduce from Proposition 4 that the expectations of $A_1, \ldots, A_4$ are also triangular fuzzy numbers, and they are given by:

$$\begin{array}{ll} E_{A_1} = (0.41, 0.51, 0.61) & E_{A_2} = (0.38, 0.47, 0.59). \\ E_{A_3} = (0.64, 0.77, 0.88) & E_{A_4} = (0.38, 0.47, 0.59). \end{array}$$

Then, applying Propositions 4 and 5, we obtain the following results:

	A_1	A_2	A_3	A_4
A_1	·	$\succeq_{\mathrm{EU}_{2,5}}$	–	$\succeq_{\mathrm{EU}_{2,5}}$
A_2	–	·	–	$\succeq_{\mathrm{EU}_{2,5}}$
A_3	$\succeq_{\mathrm{EU}_1}$	$\succeq_{\mathrm{EU}_1}$	·	$\succeq_{\mathrm{EU}_1}$
A_4	–	–	–	·

Again A_3 seems to be the most adequate option, because it is preferable to the other alternatives with respect to the first extension of the expected utility (and as a consequence also with respect to any of the other extensions).

6 Conclusions

Stochastic orders are methods for the comparison of random quantities. When the random variables to be compared are imprecisely described, they can be modeled by means of fuzzy random variables. This work presents a first approach to the extension of stochastic orders to the comparison of fuzzy random variables. We have considered two possibilities: the comparison of the images of the fuzzy random variables by means of a fuzzy ranking, and the comparison of the expectations by means of stochastic orders on possibility measures. We have investigated in more detail the particular case of fuzzy random variables whose images are triangular fuzzy numbers, and showed that the proposed methods can be simplified in that case. In addition, we have illustrated these methods in a decision making problem.

There are still several open lines of research on the comparison of fuzzy random variables. On the one hand, it is possible to extend other stochastic orders, such as stochastic dominance [11], to this context; on the other hand, we would like to deepen into the comparison of the properties of the different fuzzy rankings proposed in the literature with respect to this problem.

Finally, a different approach would be the comparison of fuzzy random variables by means of their α-cuts. In this case, the comparison is reduced to the comparison of random sets, and we can consider notions of *strong* or *weak* preference, depending on whether the comparison holds for every or any α-cut. Note also that the comparison of random sets can be made in two different ways: on the one hand, we can consider a stochastic order on random variables, and apply it to the sets of measurable selections by means of Definition 2 [9]; or we could also consider other stochastic orders for random sets, such as the ones considered in [1].

Acknowledgments. The research in this paper is partly supported by the Science and Education Ministry FPU grant AP2009-1034 and the Spanish Ministry of Science and Innovation grant MTM2010-17844.

References

1. Cascos, I., Molchanov, I.: A stochastic order for random vectors and random sets based on the Aumann expectation. Stat. Prob. Lett 63, 295–305 (2003)
2. De Schuymer, B., De Meyer, H., De Baets, B.: A fuzzy approach to stochastic dominance of random variables. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS, vol. 2715, pp. 253–260. Springer, Heidelberg (2003)
3. De Schuymer, B., De Meyer, H., De Baets, B., Jenei, S.: On the cycle-transitivity of the dice model. Theory and Decision 54, 261–285 (2003)
4. Dengjie, Z.: The limit theorems for expectation of fuzzy random variables. Int. Journal of Pure and Applied Mathematics 54(4), 489–496 (2009)
5. Dubois, D., Prade, H.: Ranking fuzzy numbers in the setting of possibility theory. Information Sciences 30, 183–224 (1983)
6. Kruse, R., Meyer, D.R.: Statistics with vague data. Reidel Publishing Company (1987)
7. Loquin, K., Dubois, D.: Kriging and epistemic uncertainty: A critical discussion. In: Jeansoulin, R., Papini, O., Prade, H., Schockaert, S. (eds.) Methods for Handling Imperfect Spatial Information. STUDFUZZ, vol. 256, pp. 269–305. Springer, Heidelberg (2010)
8. Merigó, J.M., Casanovas, M., Yang, J.-B.: Group decision making with expertons and uncertain generalized probabilistic weighted aggregation operators. European Journal of Operational Research 235, 215–224 (2014)
9. Montes, I., Miranda, E., Montes, S.: Decision making with imprecise probabilities and utilities by means of statistical preference and stochastic dominance. European Journal of Operational Research 234(1), 209–220 (2014)
10. Montes, I., Miranda, E., Montes, S.: Stochastic dominance with imprecise information. Computational Statistics and Data Analysis 71, 868–886 (2014)
11. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. Wiley (2002)
12. Puri, M.L., Ralescu, D.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)

Interval Compositional Data: Problems and Possibilities

Ondřej Pavlačka and Karel Hron

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University Olomouc,
17. listopadu 12, 771 46 Olomouc, Czech Republic
ondrej.pavlacka@upol.cz, hronk@seznam.cz

Abstract. In statistics, compositional data are defined as multivariate observations that quantitatively describe contributions of parts on a whole, carrying exclusively relative information. As a consequence, compositions can be represented as proportions or percentages without loss of information (contained in ratios between parts). Nevertheless, in the practice parts of compositional data are frequently formed by intervals; for example, concentrations of chemical elements are provided not as exact numbers, but rather in an interval range. Intuitively, a natural question arises, whether the relative information is preserved, when the original compositional data with interval-valued parts are represented in proportions. Namely, from the arithmetic properties of interval data, normalizing of intervals does not simply follow the case of real values, but a special procedure according to constrained interval arithmetic is needed. The aim of the contribution is to discuss possibilities of representing the interval compositional data in proportions.

Keywords: compositional data, Aitchison geometry, interval arithmetic, descriptive statistics.

1 Introduction

The concept of compositional data frequently occurs in many applications, covering such situations, where not the absolute values of variables, but rather relative information they contain is of primary interest [1,8]. Typical examples are formed by concentrations of chemical elements in a rock (in mg/kg), proportional representation of political parties resulting from elections, but also household expenditures on various costs (like foodstuff, housing, clothing, culture, etc.), when the relative structure of costs is to be analyzed. Consequently, compositional data are popularly represented as multivariate observations with a constant sum constraint (like in proportions or percentages). Nevertheless, the above examples clearly imply that compositions themselves are not necessarily induced by any such constraint (household expenditures can be represented both in the original units, like EUR or USD, and in proportions, the ratios between parts remain the same).

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_4

In the practice also such situations occur that compositional parts are represented by intervals rather than by precise values. One natural source of such data comes from aggregation of information over individuals in symbolic data analysis [2] in order to obtain representants of specified sets of individuals (like household expenditures in different parts of a certain region) that capture variability of the aggregation process. Another source of interval compositional data is formed by measurement process itself that leads naturally to unprecise values. Such situations arise usually in analytical chemistry or geochemistry, e.g. due rounding effects of values near to detection limit. In contrast to symbolic data analysis case, here the interval values of variables are often combined with precise ones what makes the use of procedures based on logarithmic transformation [3] conceptually not possible.

To illustrate the methodology, presented further, we use a small real-world data set, obtained from The National Institute of Public Health of Czech Republic (2011), where chemical composition of seven popular mineral waters was analyzed. For our purposes, just four chemical elements were chosen (calcium, sodium, magnesium and potassium), and the resulting values (measured in mg/l and already collected in form of interval values with lower and upper boundary) are presented in Table 1 (the mineral waters are listed with their original Czech names). We can observe that the interval values in the data set occur in all variables, except for the first mineral water (called Hanácká kyselka) and potassium in case of Magnesia.

Table 1. Interval concentrations of chemical elements (in mg/l) in Czech mineral waters

Mineral waters	Calcium		Sodium		Magnesium		Potasium	
Hanácká kyselka	275.0	275.0	68.0	68.0	251.0	251.0	17.7	17.7
Korunní	78.3	86.5	29.5	30.9	98.0	111.1	23.0	25.5
Magnesia	35.3	41.3	179.0	200.0	4.3	6.8	1.4	1.4
Mattoni	87.6	88.6	24.8	24.9	71.9	79.8	18.0	19.0
Ondrášovka	192.0	199.4	19.4	19.8	29.2	30.9	1.4	1.6
Poděbradka	142.2	145.5	45.4	49.3	344.0	360.0	47.0	49.8
Dobrá voda	8.7	10.7	9.5	9.7	9.5	10.0	9.3	9.4

Obviously, also for interval compositional data, analogously as for compositions with precise values, not absolute values of single element concentrations, but rather their relative contributions to the overall chemical composition of mineral waters is of primary interest. In other words, also here the ratios between (interval) compositional parts form the source of relevant information. Nevertheless, due to limitations of interval arithmetics, treatment of interval compositional data is more complex than in the standard (precise) case. The aim of this contribution is to draw up possible problems and challenges, related to geometrical properties and subsequent statistical analysis of interval compositional data, that might lead to a concise methodology in the future.

The paper is organized as follows. In the next section, basics of interval arithmetics are refreshed, with a focus on positive interval data (forming the compositional parts). Section 3 is devoted to the problem of forming the ratios of compositional data. In Section 4, problems related to proportional representation of interval compositional data are analyzed. Consequently, implications for descriptive statistics of interval compositions are briefly discussed. Finally, possibilities of further development are collected in the last section.

2 Computing with Intervals

Before we proceed to introduce interval compositional data and their geometrical concepts, let us briefly refresh basic possibilities of computing with intervals. Due to definition of compositional data, it is sufficient to restrict the general case for positive intervals only. We will distinguish two cases: First, we will assume that the input intervals are independent, i.e. all combinations of values belonging to the intervals are admissible. Second, we will consider interactive input intervals, where the set of all admissible combinations of values is given.

Let $I_1, \dots, I_n$ be independent intervals and $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. Then

$$f(I_1, \dots, I_n) = \left[\underline{y}, \overline{y}\right],$$

where

$$\underline{y} = \min\{f(x_1, \dots, x_n) \mid x_i \in I_i,\ i = 1, \dots, n\},$$
$$\overline{y} = \max\{f(x_1, \dots, x_n) \mid x_i \in I_i,\ i = 1, \dots, n\}.$$

If f stands for one of the basic arithmetic operations, we get the well-known concept of *standard interval arithmetic*. Let $[a, b]$, $0 < a \leq b$, and $[c, d]$, $0 < c \leq d$, be independent intervals. Then arithmetic operations are extended as follows:

$$\begin{aligned} [a, b] + [c, d] &= [a + c, b + d]\,, \\ [a, b] - [c, d] &= [a - d, b - c]\,, \\ [a, b] \cdot [c, d] &= [a \cdot c, b \cdot d]\,, \\ \frac{[a, b]}{[c, d]} &= \left[\frac{a}{d}, \frac{b}{c}\right]. \end{aligned}$$

However, the above concept cannot be applied in the case when it is given a constraint set $Q \subset \mathbb{R}^n$ that represents all admissible combinations of the values of $x_1, \dots, x_n$ (see e.g. [5]). If $Q \cap (I_1 \times \ldots \times I_n)$ is a nonempty convex set, then

$$f(I_1, \dots, I_n; Q) = \left[\underline{y}, \overline{y}\right],$$

where

$$\underline{y} = \min\{f(x_1, \dots, x_n) \mid x_i \in I_i,\ i = 1, \dots, n,\ (x_1, \dots, x_n) \in Q\}$$
$$\overline{y} = \max\{f(x_1, \dots, x_n) \mid x_i \in I_i,\ i = 1, \dots, n,\ (x_1, \dots, x_n) \in Q\}.$$

In our case, the role of a constraint set Q will play, for instance, the set representing proportional representation of interval compositional data.

3 Ratios of Compositional Parts

Following the Introduction section, a sample of D-part compositional data are positive vectors $\mathbf{x}_i := (x_{i1}, \dots, x_{iD})$, $i = 1, \dots, n$, that describe quantitatively contributions of parts on a whole, carrying exclusively relative information. This means that the relevant information is expressed by the ratios $r^i_{jk} := x_{ij}/x_{ik}$, $j, k = 1, \dots, D$, $j \neq k$.

Now, let us consider the case of interval compositional data

$$\mathbf{X}_i := \left([\underline{x}_{i1}, \overline{x}_{i1}], \dots, [\underline{x}_{iD}, \overline{x}_{iD}]\right), \quad i = 1, \dots, n,$$

where

$$0 < \underline{x}_{ij} \leq \overline{x}_{ij},\ j = 1, \dots, D.$$

For the sake of simplicity, let us assume further that $[\underline{x}_{ij}, \overline{x}_{ij}]$ and $[\underline{x}_{lk}, \overline{x}_{lk}]$ are independent intervals for any $i, l = 1, \dots, n$, and $j, k = 1, \dots, D$, $j \neq k$.

According to the assumption, we can apply the concept of standard interval arithmetics for computing the ratios between the compositional parts:

$$R^i_{jk} := \frac{[\underline{x}_{ij}, \overline{x}_{ij}]}{[\underline{x}_{ik}, \overline{x}_{ik}]} = \left[\frac{\underline{x}_{ij}}{\overline{x}_{ik}}, \frac{\overline{x}_{ij}}{\underline{x}_{ik}}\right], \quad i = 1, \dots, n,\ j, k = 1, \dots, D,\ j \neq k. \tag{1}$$

For illustration, the ratios between concentrations of chemical elements presented in Table 1 are shown in Table 2.

Table 2. Ratios between concentrations of chemical elements in Czech mineral waters

Mineral waters	Calcium/Sodium		Calcium/Magnesium		Calcium/Potasium	
Hanácká kyselka	4.044	4.044	1.096	1.096	15.537	15.537
Korunní	2.534	2.932	0.705	0.883	3.071	3.761
Magnesia	0.177	0.231	5.191	9.605	25.214	29.500
Mattoni	3.518	3.573	1.098	1.232	4.611	4.922
Ondrášovka	9.697	10.278	6.214	6.829	120.0	142.429
Poděbradka	2.884	3.205	0.395	0.4236	2.855	3.096
Dobrá voda	0.897	1.126	0.870	1.126	0.926	1.151

For possible further dealing with interval ratios R^i_{jk} obtained by (1), it is worth to note that R^i_{jk} and R^i_{jl}, $k \neq l$, are not independent intervals since the same interval $[\underline{x}_{ij}, \overline{x}_{ij}]$ is used for their calculation.

4 Proportional Representation

The original compositions $\mathbf{x}_i$, $i = 1, \dots, n$, are often represented so that the sums of the components for each composition are equal to an arbitrary (but fixed) constant $\kappa > 0$. Such a representation is formally expressed by the *closure operation*

$$\mathcal{C}(\mathbf{x}_i) := \left(\frac{\kappa \cdot x_{i1}}{\sum_{j=1}^{D} x_{ij}}, \ldots, \frac{\kappa \cdot x_{iD}}{\sum_{j=1}^{D} x_{ij}} \right), \quad i = 1, \ldots, n.$$

The constant κ is popularly taken as 1 (or 100) in case of proportional (percentage) representation. It is essential that the proportional representation keeps the ratios between the compositional parts as

$$\frac{\frac{\kappa \cdot x_{ik}}{\sum_{j=1}^{D} x_{ij}}}{\frac{\kappa \cdot x_{il}}{\sum_{j=1}^{D} x_{ij}}} = \frac{x_{ik}}{x_{il}}, \quad k, l = 1, \ldots, D.$$

Note that the resulting *scale invariance* is one of the basic properties of compositional data, reflected also by the Aitchison geometry [8] that forms a natural algebraic-geometrical structure of compositions. Without the loss of generality, we will assume $\kappa = 1$ further in the paper.

For the interval compositional data $\mathbf{X}_i$, $i = 1, \ldots, n$, the situation becomes more complex. Observe that in the proportions

$$\mathcal{C}(\mathbf{x}_i)_k := \frac{x_{ik}}{\sum_{j=1}^{D} x_{ij}}, \ k = 1, \ldots, D,$$

the variable x_{ik} appears both in the numerator and the denominator. Hence, we cannot apply the concept of standard interval arithmetic and compute the k-th interval proportion in the following way:

$$\frac{[\underline{x}_{ik}, \overline{x}_{ik}]}{\sum_{j=1}^{D} [\underline{x}_{ij}, \overline{x}_{ij}]} = \frac{[\underline{x}_{ik}, \overline{x}_{ik}]}{\left[\sum_{j=1}^{D} \underline{x}_{ij}, \sum_{j=1}^{D} \overline{x}_{ij}\right]} = \left[\frac{\underline{x}_{ik}}{\sum_{j=1}^{D} \overline{x}_{ij}}, \frac{\overline{x}_{ik}}{\sum_{j=1}^{D} \underline{x}_{ij}} \right],$$

as the numerator and the denominator are not independent intervals. The correct procedure for computing the intervals $[\underline{c}_{ik}, \overline{c}_{ik}]$, $k = 1, \ldots, D$, that express the ranges of particular proportions from $\mathcal{C}(\mathbf{X}_i)$ is given in the following way (the formulas were developed for the first time in [4] for normalizing interval weights):

$$\underline{c}_{ik} = \min \left\{ \frac{x_{ik}}{\sum_{j=1}^{D} x_{ij}} \mid x_{ij} \in \left[\underline{x}_{ij}, \overline{x}_{ij}\right], \ j = 1, \ldots, D \right\} = \frac{\underline{x}_{ik}}{\underline{x}_{ik} + \sum_{j=1,\, j \neq k}^{D} \overline{x}_{ij}},$$

$$\overline{c}_{ik} = \max \left\{ \frac{x_{ik}}{\sum_{j=1}^{D} x_{ij}} \mid x_{ij} \in \left[\underline{x}_{ij}, \overline{x}_{ij}\right], \ j = 1, \ldots, D \right\} = \frac{\overline{x}_{ik}}{\overline{x}_{ik} + \sum_{j=1,\, j \neq k}^{D} \underline{x}_{ij}}.$$

The interval proportions of concentrations of chemical elements presented in Table 1 are given in Table 3.

However, the obtained intervals $[\underline{c}_{ik}, \overline{c}_{ik}]$, $k = 1, \ldots, D$, are not independent, so it is not correct to compute their ratios by means of standard interval arithmetic. Applying the results concerning normalization of interval weights that were proved in [7], we find out that the following general relations hold:

Table 3. Interval proportions of concentrations of chemical elements from Table 1

Mineral waters	Calcium		Sodium		Magnesium		Potasium	
Hanácká kyselka	0.45	0.45	0.111	0.111	0.41	0.41	0.029	0.029
Korunní	0.319	0.365	0.117	0.134	0.407	0.459	0.092	0.011
Magnesia	0.145	0.183	0.783	0.823	0.017	0.031	0.006	0.006
Mattoni	0.415	0.436	0.117	0.123	0.352	0.38	0.085	0.094
Ondrášovka	0.786	0.8	0.077	0.082	0.117	0.127	0.006	0.007
Poděbradka	0.237	0.25	0.076	0.085	0.584	0.605	0.078	0.086
Dobrá voda	0.23	0.274	0.24	0.261	0.242	0.267	0.234	0.253

$$R^i_{jk} \subseteq \frac{[\underline{c}_{ij},\overline{c}_{ij}]}{[\underline{c}_{ik},\overline{c}_{ik}]} \quad j,k=1,\ldots,D,\ j\neq k.$$

Example 1. Let us consider the interval ratio between concentrations of calcium and sodium in mineral water Korunní $[2.534, 2.932]$ (see Table 2). If we compute, by means of the standard interval arithmetic operations, the ratio between interval proportions of calcium and sodium on the whole presented in Table 3, we obtain the following result:

$$\frac{[0.319, 0.365]}{[0.117, 0.134]} = [2.373, 3.125]\,.$$

We can see that the interval ratio $[2.534, 2.932]$ is indeed a strict subset of $[2.373, 3.125]$.

The interactions among the proportions $[\underline{c}_{ik},\overline{c}_{ik}]$, $k = 1,\ldots,D$, mean that the proper proportional representation of interval compositional data $\mathbf{X}_i$, $i = 1,\ldots,n$, has to be given in the following way:

$$\mathcal{C}(\mathbf{X}_i) := \left\{\mathcal{C}(\mathbf{x}_i) \in [0,1]^D \mid \mathbf{x}_i \in [\underline{x}_{i1},\overline{x}_{i1}] \times \ldots \times [\underline{x}_{iD},\overline{x}_{iD}]\right\}. \qquad (2)$$

Employing the results proved in [6] concerning normalization of interval weights, we can see that, unless $D = 2$, the interval proportions $[\underline{c}_{ik},\overline{c}_{ik}]$, $k = 1,\ldots,D$, alone do not carry the whole information about the proportional representation of interval compositional data. From (2), we can see that we still have to know the initial interval compositional data $\mathbf{X}_i$, $i = 1,\ldots,n$. For $D = 2$, it is on the contrary sufficient to know only one interval proportion, e.g. $[\underline{c}_{i1},\overline{c}_{i1}]$, $\mathcal{C}(\mathbf{X}_i)$ can be then given as follows:

$$\mathcal{C}(\mathbf{X}_i) = \left\{(c_{i1}, c_{i2}) \in [0,1]^2 \mid c_{i1} \in [\underline{c}_{i1},\overline{c}_{i1}]\,,\ c_{i2} = 1 - c_{i1}\right\}.$$

Remark 1. Note that if the ratios between two proportions are calculated properly, they are equal to the ratios between the corresponding original compositional parts (the following procedure is inspired by the procedure introduced in [7] for computing the ratios between normalized fuzzy weights). For $j,k = 1,\dots,D$, $j \neq k$, let us denote the proper ratio between the j-th and k-th proportions by $[\underline{r}^i_{jk}, \overline{r}^i_{jk}]$. Then

$$\begin{aligned}\underline{r}^i_{jk} &= \min\Big\{\tfrac{c_{ij}}{c_{ik}} \mid c_{ij} \text{ and } c_{ik} \text{ express the } j\text{-th and } k\text{-th components} \\ &\qquad\qquad \text{of at least one } \mathcal{C}(\mathbf{x}_i) \in \mathcal{C}(\mathbf{X}_i)\Big\},\\ \overline{r}^i_{jk} &= \max\Big\{\tfrac{c_{ij}}{c_{ik}} \mid c_{ij} \text{ and } c_{ik} \text{ express the } j\text{-th and } k\text{-th components} \\ &\qquad\qquad \text{of at least one } \mathcal{C}(\mathbf{x}_i) \in \mathcal{C}(\mathbf{X}_i)\Big\}.\end{aligned}$$

It can be shown (see [7, Theorem 8]) that $[\underline{r}^i_{jk}, \overline{r}^i_{jk}] = R^i_{jk}$ for any $j,k = 1,\dots,D$, $j \neq k$.

5 Descriptive Statistics

Specific properties of (precise) compositional data, captured by the Aitchison geometry, should be reflected also by their descriptive statistics [1,9]. For instance, the arithmetic mean as a measure of location needs to be replaced by the geometric mean (centre) of compositional data, $\mathbf{g}(\mathbf{x}) := \left(g\left(\mathbf{x}^1\right),\dots,g\left(\mathbf{x}^D\right)\right)$, where $\mathbf{x}^j := (x_{1j},\dots,x_{nj})$ and $g\left(\mathbf{x}^j\right) := \sqrt[n]{\prod_{i=1}^n x_{ij}}$, $j = 1,\dots,D$. Note that the centre can be computed from an arbitrary representation of the input compositions $\mathbf{x}_1,\dots,\mathbf{x}_n$, the ratios between parts of $\mathbf{g}(\mathcal{C}(\mathbf{x}))$ remain always the same, i.e.

$$\frac{g\left(\mathbf{x}^j\right)}{g\left(\mathbf{x}^k\right)} = \frac{g\left(\mathcal{C}\left(\mathbf{x}^j\right)\right)}{g\left(\mathcal{C}\left(\mathbf{x}^k\right)\right)}, \quad j,k = 1,\dots,D. \tag{3}$$

Now, let us consider the case of interval compositional data $\mathbf{X}_i$, $i = 1,\dots,n$, introduced in Section 3. Since the particular intervals are assumed to be independent, the centre of these data is given as a vector $\mathbf{g}(\mathbf{X}) = \left(g\left(\mathbf{X}^1\right),\dots,g\left(\mathbf{X}^D\right)\right)$, where

$$\mathbf{X}^j := \left(\left[\underline{x}_{1j}, \overline{x}_{1j}\right],\dots,\left[\underline{x}_{nj}, \overline{x}_{nj}\right]\right), \quad j = 1,\dots,D,$$

and

$$g\left(\mathbf{X}^j\right) := \left[\sqrt[n]{\prod\nolimits_{i=1}^n \underline{x}_{ij}}, \sqrt[n]{\prod\nolimits_{i=1}^n \overline{x}_{ij}}\right], \quad j = 1,\dots,D.$$

Note that the particular intervals $g\left(\mathbf{X}^j\right)$, $j = 1,\dots,D$, are independent. Hence, their ratios have to be computed applying the concept of standard interval arithmetic.

At the end of this section, let us verify the validity of equality (3) in the case of interval compositional data. Let

$$\mathcal{C}\left(\mathbf{X}^j\right) := \left(\left[\underline{c}_{1j}, \overline{c}_{1j}\right],\dots,\left[\underline{c}_{nj}, \overline{c}_{nj}\right]\right), \quad j = 1,\dots,D,$$

and $\mathbf{g}(\mathcal{C}(\mathbf{X})) = \left(g\left(\mathcal{C}\left(\mathbf{X}^1\right)\right), \ldots, g\left(\mathcal{C}\left(\mathbf{X}^D\right)\right)\right)$, where, for $i = 1, \ldots, n$, $j = 1, \ldots, D$, $\left[\underline{c}_{ij}, \overline{c}_{ij}\right]$ expresses the range of proportion of the j-th part in $\mathbf{X}_i$. Since the intervals $\left[\underline{c}_{ij}, \overline{c}_{ij}\right]$, $i = 1, \ldots, n$, $j = 1, \ldots, D$, are not independent, also the obtained intervals $g\left(\mathcal{C}\left(\mathbf{X}^1\right)\right), \ldots, g\left(\mathcal{C}\left(\mathbf{X}^D\right)\right)$ are not independent and their ratios cannot be computed applying the concept of standard interval arithmetic. If we do so, we obtain the following relation instead of equality (3):

$$\frac{g\left(\mathbf{X}^j\right)}{g\left(\mathbf{X}^k\right)} \subseteq \frac{g\left(\mathcal{C}\left(\mathbf{X}^j\right)\right)}{g\left(\mathcal{C}\left(\mathbf{X}^k\right)\right)}, \quad j, k = 1, \ldots, D.$$

Therefore, for retaining the validity of equality (3), we have to respect the interactions among $g\left(\mathcal{C}\left(\mathbf{X}^1\right)\right), \ldots, g\left(\mathcal{C}\left(\mathbf{X}^D\right)\right)$ when computing their ratios.

6 Future Work

Interval compositional data form a natural extension of the precise case. We have studied possible ways of dealing with such data. Future work will be aimed at extension of other procedures developed for dealing with compositional data. Another problem worth to study will be utilization of the information about precise sum of compositional parts, that is frequently available in the practice.

Acknowledgments. The paper is supported by the Operational Program Education for Competitiveness-European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic), and the grant IGA_PrF_2014028 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London (1986)
2. Billard, L., Diday, E.: From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. J. Amer. Statist. Assoc. 98, 470–487 (2003)
3. Brito, P., Silva, A.P.D.: Modelling Interval Data with Normal and Skew-Normal Distributions. J. Appl. Stat. 39, 3–20 (2012)
4. Dubois, D., Prade, H.: The Use of Fuzzy Numbers in Decision Analysis. In: Gupta, M.M., Sanchez, E. (eds.) Fuzzy Information and Decision Processes, pp. 309–321. North-Holland, Amsterdam (1982)
5. Klir, G.J., Pan, Y.: Constrained Fuzzy Arithmetic: Basic Questions and Some Answers. Soft Comput. 2, 100–108 (1998)
6. Pavlačka, O.: Note on the lack of equality between fuzzy weighted average and fuzzy convex sum. Fuzzy Set. Syst. 213, 102–105 (2013)
7. Pavlačka, O.: On various approaches to normalization of interval and fuzzy weights. Fuzzy Set. Syst. 243, 110–130 (2014)
8. Pawlowsky-Glahn, V., Buccanti, A. (eds.): Compositional Data Analysis: Theory and Applications. Wiley, Chichester (2011)
9. Pawlowsky-Glahn, V., Egozcue, J.J.: BLU Estimators and Compositional Data. Math. Geol. 34, 259–274 (2002)

A New Definition of Evaluation/Defuzzification of an Interval Type-2 Fuzzy Set

Luca Anzilli[1], Gisella Facchinetti[1], and Tommaso Pirotti[2]

[1] Department of Management, Economics, Mathematics and Statistics,
University of Salento, Italy
{luca.anzilli,gisella.facchinetti}@unisalento.it
[2] Department of Economics "Marco Biagi",
University of Modena and Reggio Emilia, Italy
tommaso.pirotti@unimore.it

Abstract. In this paper we propose a new evaluation/defuzzification formula for an Interval Type-2 Fuzzy Quantity (IT2 FQ), that is an Interval Type-2 Fuzzy Set (IT2 FS) defined by two Type-1 Fuzzy Quantities (T1 FQs) having membership functions that may be neither convex nor normal. We start from a parametric formula to evaluate them and we propose to call the IT2 FQ value their average. To compare the results we obtain changing the parameters, we use the final output of an example of Interval Type-2 Fuzzy Logic System (IT2 FLS).

Keywords: Fuzzy sets, fuzzy quantities, interval type-2 fuzzy sets, evaluation.

1 Introduction

Type-2 fuzzy sets and systems generalize type-1 fuzzy sets and systems so that more uncertainty can be handled. When fuzzy sets enter in scientific world, one of critics is due to the fact that the membership function of a Type-1 Fuzzy Set (T1 FS) has no uncertainty associated with it. This fact seems to contradict the word "fuzzy". In 1975 Prof. Lotfi A. Zadeh [20] proposed more sophisticated kinds of fuzzy sets, he called Type-2 Fuzzy Sets (T2 FSs). A T2 FS lets us incorporate uncertainty about the membership function into fuzzy set theory, and is a way to address the above criticism of T1 FS heads-on. The membership function of a T2 FS is three-dimensional, where the third dimension is the value of the membership function at each point on its two-dimensional domain which is called its footprint of uncertainty (FOU). Interval Type-2 Fuzzy Sets (IT2 FSs) are particular T2 FSs in which third dimension value is constant (e.g., 1). This means that no new information is contained in the third dimension of an IT2 FS and only the FOU is used to describe it. An IT2 FS is completely described by two T1 FSs whose membership functions are the lower and upper bounds of its FOU.

After the wide number of applications of Type-1 Fuzzy Logic Systems (T1 FLSs), even the Interval Type-2 Fuzzy Logic Systems (IT2 FLSs) started and

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_5

found a lot of interesting and successful applications in signal processing, fingerprints detection and in Computing With Words fields. The researches on IT2 FLSs had a wide impulse by Prof. Jerry Mendel and others researchers works [12,13,14,15]. The final output of an IT2 FLS is an IT2 FS and thus one needs methods for the evaluation/defuzzification of an IT2 FS. Karnik and Mendel [11] proposed a defuzzification method based on an algorithm that evaluates an IT2 FS taking the average of the centroids of T1 FSs embedded in the FOU zone.

This paper goes in the same direction and proposes a parametric evaluation/defuzzification formula for an Interval Type-2 Fuzzy Quantity (IT2 FQ), that is an IT2 FS defined by two Type-1 Fuzzy Quantities (T1 FQs) whose membership functions may be neither convex nor normal. We start from a parametric formula for the evaluation of the two T1 FQs and we propose to call the IT2 FQ value their average. This approach allows us, by changing the set of parameters, to recover the T1 FQs evaluations proposed by Fortemps and Roubens [8,3], Yager and Filev [18,19], Anzilli and Facchinetti [3] and Center of Gravity (COG). To illustrate how our method works, we apply it to the final output of an example of IT2 FLS and compare the numerical results we obtain changing the set of parameters. In Section 2 and Section 3 we introduce the concepts of IT2 FS and IT2 FQ. In Section 4 we give an example of IT2 FLS and in section 5 we present the evaluation model for an IT2 FQ and apply it to the defuzzification of the final output of the IT2 FLS.

2 Interval Type-2 Fuzzy Sets

We give a short presentation of T2 FSs and IT2 FSs (for detail see [15]).

Definition 1. *A T2 FS $\tilde{A}$ in the universe of discourse X is characterized by a type-2 membership function $\mu_{\tilde{A}}(x,u)$ where $x \in X$ and $u \in J_x \subseteq [0,1]$, i.e.*

$$\tilde{A} = \{((x,u), \mu_{\tilde{A}}(x,u))\,;\ \forall x \in X\,, \forall u \in J_x \subseteq [0,1]\}$$

in which $0 \leq \mu_{\tilde{A}}(x,u) \leq 1$. J_x is a closed interval of real numbers. A T2 FS $\tilde{A}$ can also be represented as $\tilde{A} = \int_{x\in X}\int_{u\in J_x} \mu_{\tilde{A}}(x,u)/(x,u)$.

Definition 2. *If all $\mu_{\tilde{A}}(x,u) = 1$ then $\tilde{A}$ is called an IT2 FS.*

An IT2 FS $\tilde{A}$ can be considered as a special case of a T2 FS and it can be expressed as $\tilde{A} = \int_{x\in X}\int_{u\in J_x} 1/(x,u)$. J_x is called the primary membership of x.

The footprint of uncertainty (FOU) of an IT2 FS $\tilde{A}$ is defined by FOU$(\tilde{A}) = \bigcup_{x\in X} J_x$. The FOU is a complete description of an IT2 FS. The upper membership function $\mu^U_{\tilde{A}}$ and the lower membership function $\mu^L_{\tilde{A}}$ of an IT2 FS $\tilde{A}$ are defined as the two type-1 membership functions that bound the FOU. Thus $J_x = [\mu^L_{\tilde{A}}(x), \mu^U_{\tilde{A}}(x)]$ for all $x \in X$. In the following an IT2 FS $\tilde{A}$ will be denoted by $\tilde{A} = (A^L, A^U)$, where A^L and A^U are the T1 FSs with membership functions $\mu_{A^L} = \mu^L_{\tilde{A}}$ and $\mu_{A^U} = \mu^U_{\tilde{A}}$, respectively.

The intersection and the union of two IT2 FSs $\tilde{A}$, $\tilde{B}$ are defined as the IT2 FSs given by

$$\tilde{A} \sqcap \tilde{B} = \int_{x \in X} \int_{u \in [\mu^L_{\tilde{A}\sqcap\tilde{B}}(x), \mu^U_{\tilde{A}\sqcap\tilde{B}}(x)]} 1/(x,u)$$

$$\tilde{A} \sqcup \tilde{B} = \int_{x \in X} \int_{u \in [\mu^L_{\tilde{A}\sqcup\tilde{B}}(x), \mu^U_{\tilde{A}\sqcup\tilde{B}}(x)]} 1/(x,u)$$

with $\mu^L_{\tilde{A}\sqcap\tilde{B}}(x) = T(\mu^L_{\tilde{A}}(x), \mu^L_{\tilde{B}}(x))$, $\mu^U_{\tilde{A}\sqcap\tilde{B}}(x) = T(\mu^U_{\tilde{A}}(x), \mu^U_{\tilde{B}}(x))$, $\mu^L_{\tilde{A}\sqcup\tilde{B}}(x) = S(\mu^L_{\tilde{A}}(x), \mu^L_{\tilde{B}}(x))$ and $\mu^U_{\tilde{A}\sqcup\tilde{B}}(x) = S\mu^U_{\tilde{A}}(x), \mu^U_{\tilde{B}}(x))$, where T is the t-norm operator and S is the t-conorm operator.

3 Interval Type-2 Fuzzy Quantities

We now introduce the concept of T1 FQ (see [3,4]) and the definition of IT2 FQ.

Definition 3. *Let N be a positive integer and let $a_1, a_2, \ldots, a_{4N}$ be real numbers with $a_1 < a_2 \leq a_3 < a_4 \leq a_5 < a_6 \leq a_7 < a_8 \leq a_9 < \cdots < a_{4N-2} \leq a_{4N-1} < a_{4N}$. We call type-1 fuzzy quantity*

$$A = (a_1, a_2, \ldots, a_{4N};\ h_1, h_2, \ldots, h_N,\ h_{1,2}, h_{2,3}, \ldots, h_{N-1,N}) \tag{1}$$

where $0 < h_j \leq 1$ for $j = 1, \ldots, N$ and $0 \leq h_{j,j+1} < \min\{h_j, h_{j+1}\}$ for $j = 1, \ldots, N-1$, the fuzzy set defined by a continuous membership function $\mu : \mathbb{R} \to [0,1]$, with $\mu(x) = 0$ for $x \leq a_1$ or $x \geq a_{4N}$, such that for $j = 1, 2, \ldots, N$

(i) μ is strictly increasing in $[a_{4j-3}, a_{4j-2}]$, with $\mu(a_{4j-3}) = h_{j-1,j}$ and $\mu(a_{4j-2}) = h_j$,
(ii) μ is constant in $[a_{4j-2}, a_{4j-1}]$, with $\mu \equiv h_j$,
(iii) μ is strictly decreasing in $[a_{4j-1}, a_{4j}]$, with $\mu(a_{4j-1}) = h_j$ and $\mu(a_{4j}) = h_{j,j+1}$,

and for $j = 1, 2, \ldots, N-1$

(iv) μ is constant in $[a_{4j}, a_{4j+1}]$, with $\mu \equiv h_{j,j+1}$,

where $h_{0,1} = h_{N,N+1} = 0$. Thus the height of A is $h_A = \max_{j=1,\ldots,N} h_j$.

We observe that in the case $N = 1$ the T1 FQ defined in (1) is fuzzy convex, that is every α-cut A_α is a closed interval. If $N \geq 2$ the T1 FQ defined in (1) is a non-convex fuzzy set with N humps and height $h_A = \max_{j=1,\ldots,N} h_j$.

Definition 4. *We call Interval Type-2 Fuzzy Quantity (IT2 FQ) an IT2 FS $\tilde{A}$ such that $\mu^L_{\tilde{A}}$ and $\mu^U_{\tilde{A}}$ are membership functions of T1 FQs.*

If $\tilde{A}$ is an IT2 FQ we denote by A^L the T1 FQ with membership function $\mu_{A^L} = \mu^L_{\tilde{A}}$ and by A^U the T1 FQ with membership function $\mu_{A^U} = \mu^U_{\tilde{A}}$ (see Figure 2). In the following an IT2 FQ $\tilde{A}$ will be denoted by $\tilde{A} = (A^L, A^U)$.

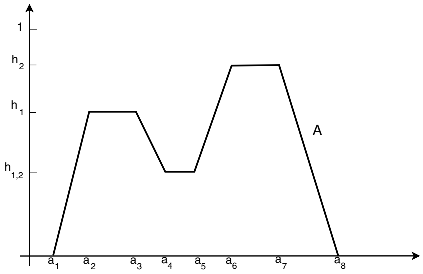


Fig. 1. Piecewise linear T1 FQ ($N = 2$)

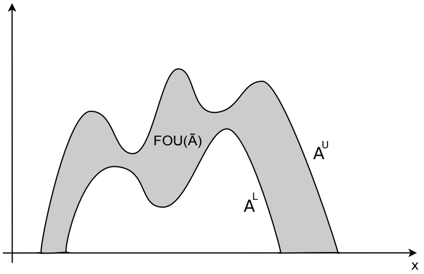


Fig. 2. IT2 FQ $\tilde{A} = (A^L, A^U)$

4 An Example of Interval Type-2 Fuzzy Logic Systems

Suppose we have an Interval Type-2 Fuzzy Logic Systems (IT2 FLS) with p inputs, $x_1, \ldots, x_p$ and one output y. Consider its rule-block characterized by M rules where the m-th rule has the form

$$R_m: \quad \text{IF } x_1 \text{ is } \tilde{F}_{1m} \text{ and } \ldots \text{ and } x_p \text{ is } \tilde{F}_{pm} \text{ THEN } y \text{ is } \tilde{G}_m \qquad m = 1, \ldots, M$$

where $\tilde{F}_{im}$, $\tilde{G}_m$ are IT2 FSs. Note that $\tilde{F}_{im}$ is the linguistic label associated with i-th antecedent in the m-th rule and $\tilde{G}_m$ is the linguistic label associated with the output variable in the m-th rule. Let us define $\tilde{F}_m = \prod_{i=1}^{p} \tilde{F}_{im}$, $m = 1, \ldots, M$. The output $\tilde{G}_m^*$ of each rule is the IT2 FS given by $\tilde{G}_m^* = \tilde{F}_m' \circ (\tilde{F}_m \to \tilde{G}_m)$, where $\circ$ is the sup-star composition operator. The final output $\tilde{G}^*$ is the IT2 FS obtained as $\tilde{G}^* = \bigsqcup_{m=1}^{M} \tilde{G}_m^*$.

We consider a singleton IT2 FLS, that is a IT2 FLS with crisp input $x' = (x_1', \ldots, x_p')$. We assume that Mamdani implications are used, $T = \min$, $S = \max$. For each rule $m = 1, \ldots, M$ we compute the firing interval $[\mu_{\tilde{F}_m}^L(x'), \mu_{\tilde{F}_m}^U(x')]$ as

$$\mu_{\tilde{F}_m}^L(x') = \min_{i=1,\ldots,p} \mu_{\tilde{F}_{im}}^L(x_i'), \qquad \mu_{\tilde{F}_m}^U(x') = \min_{i=1,\ldots,p} \mu_{\tilde{F}_{im}}^U(x_i').$$

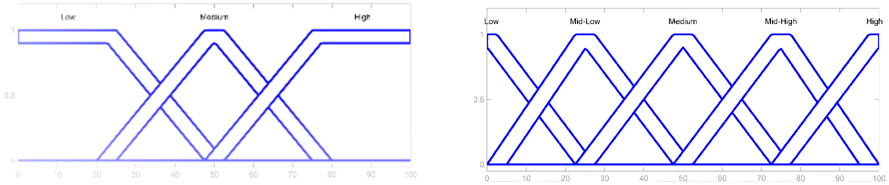
For $m = 1, \ldots, M$ the output IT2 FS of rule m, $\tilde{G}_m^* = (\tilde{G}_m^{*\,L}, \tilde{G}_m^{*\,U})$, is calculated as

$$\mu_{\tilde{G}_m^*}^L(y) = \min\left\{\mu_{\tilde{F}_m}^L(x'), \mu_{\tilde{G}_m}^L(y)\right\}, \qquad \mu_{\tilde{G}_m^*}^U(y) = \min\left\{\mu_{\tilde{F}_m}^U(x'), \mu_{\tilde{G}_m}^U(y)\right\}.$$

The final output IT2 FS $\tilde{G}^* = (\tilde{G}^{*L}, \tilde{G}^{*U})$ is obtained as $\tilde{G}^* = \bigsqcup_{m=1}^{M} \tilde{G}_m^*$, that is

$$\mu_{\tilde{G}^*}^L(y) = \max_{m=1,\ldots M} \mu_{\tilde{G}_m^*}^L(y), \qquad \mu_{\tilde{G}^*}^U(y) = \max_{m=1,\ldots M} \mu_{\tilde{G}_m^*}^U(y).$$

To show our defuzzification method we consider a very simple example of IT2 FLS with two inputs and one output. The example is the type-2 translation of a client financial risk tolerance model illustrated in [5, p.130], with a little difference on output granularity. "Financial service institutions face a difficult task in evaluating clients risk tolerance. It is a major component for the design of an investment policy and understanding the implication of possible investment options in terms of safety and suitability. Here we present a simple model of client's risk tolerance ability (RT), which depends on his/hers annual income (AI) and total net worth (TNW)". Suppose the financial experts agree to describe the input variables AI and TN by the linguistic terms {L (Low), M (Medium), H (High)} and the output variable RT by the linguistic terms {L (Low), LM (Low-Medium), M (Medium), MH (Medium-High), H (High)}. Each granule is an IT2 FS in which the domains are : $U_1 = \{x \times 10^3\,; 0 \leq x \leq 100\}$ for input AI, $U_2 = \{y \times 10^4\,; 0 \leq y \leq 100\}$ for input TN and $U_3 = \{z\,; 0 \leq z \leq 100\}$ for output RT. The real numbers x and y represent euros in thousands and hundred of thousands, correspondingly, while z takes values on a psychometric scale from 0 to 100 measuring risk tolerance. All the granules are described by triangular or trapezoidal IT2 FSs, as shown in Fig. 3.



(a) Input variables AI and TN

(b) Output variable RT

Fig. 3. Input and output variables of IT2 FLS

We assume that the financial experts selected the rules:

R_1: IF AI is L and TN is L THEN RT is L
R_2: IF AI is L and TN is M THEN RT is ML
R_3: IF AI is L and TN is H THEN RT is ML
R_4: IF AI is M and TN is L THEN RT is ML
R_5: IF AI is M and TN is M THEN RT is M
R_6: IF AI is M and TN is H THEN RT is MH
R_7: IF AI is H and TN is L THEN RT is MH
R_8: IF AI is H and TN is M THEN RT is MH
R_9: IF AI is H and TN is H THEN RT is H

If we set crisp inputs by $x = 38$ and $y = 70$, the final output is the IT2 FQ $\tilde{G}^* = (\tilde{G}^{*L}, \tilde{G}^{*U})$ shown in Fig. 4, where $\tilde{G}^{*L}$ and $\tilde{G}^{*U}$ are T1 FQs (see (1)) given by

$$\begin{aligned} G^{*L} &= (5.00, 11.91, 39.73, 43.00, 57.00, 64.20, 84.60, 95.00; 0.31, 0.47, 0.18) \\ G^{*U} &= (0.00, 11.86, 39.32, 43.41, 56.59, 63.86, 85.27, 100.00; 0.53, 0.65, 0.36)\,. \end{aligned} \tag{2}$$

5 Evaluation of Interval Type-2 Fuzzy Quantities

In [4] we propose a way to approximate a T1 FQ by an interval. Our proposal starts from Grzegorzewski's papers in which the author defines and finds the approximating interval of a fuzzy number. Starting from a distance between two fuzzy numbers and observing that any closed interval is a fuzzy number, the author defines the approximating interval of a fuzzy number as the interval of minimum distance. The distance he uses is based on the distance between intervals introduced by Trutschnig et al. [16]. This idea needs that each α-cut is an interval, that is the fuzzy set has to be convex. Hence, we cannot follow the same approach for non convex fuzzy quantities. To overcome this obstacle we noticed that Grzegorzewski's procedure may be regarded as the study of the minimum of the variance between the α-cuts family identifying a fuzzy number and a generic interval. This new way to look at the problem may be useful for non convex fuzzy quantities too.

Proposition 1. *Let A be the T1 fuzzy quantity defined in* (1) *with height h_A. Then for each $\alpha \in [0, h_A]$ there exist an integer n_α, with $1 \leq n_\alpha \leq N$, and $A_1^\alpha, \dots, A_{n_\alpha}^\alpha$ disjoint closed intervals such that $A_\alpha = \bigcup_{i=1}^{n_\alpha} A_i^\alpha = \bigcup_{i=1}^{n_\alpha} [a_i^L(\alpha), a_i^R(\alpha)]$, where we have denoted $A_i^\alpha = [a_i^L(\alpha), a_i^R(\alpha)]$, with $A_i^\alpha < A_{i+1}^\alpha$ (that is $a_i^R(\alpha) < a_{i+1}^L(\alpha)$). Thus n_α is the number of intervals producing the α-cut A_α.*

From decomposition theorem for T1 FSs and using previous result, we get

$$A = \bigcup_{\alpha \in [0,h_A]} \alpha\, A_\alpha = \bigcup_{\alpha \in [0,h_A]} \alpha \bigcup_{i=1}^{n_\alpha} A_i^\alpha = \bigcup_{\alpha \in [0,h_A]} \bigcup_{i=1}^{n_\alpha} \alpha\, A_i^\alpha$$

and thus the T1 FQ is identified by the intervals $\{A_i^\alpha;\ i = 1 \dots, n_\alpha,\ 0 \leq \alpha \leq h_A\}$.

Definition 5. *We say that $C^*(A) = [c_L^*, c_R^*]$ is an approximation interval of the T1 FQ A with respect to p, f, θ if it minimizes the weighted mean of the squared distances*

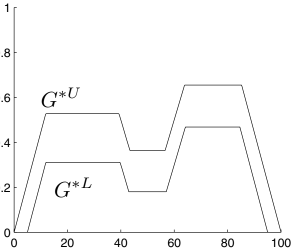


Fig. 4. Output IT2 FQ $\tilde{G}^* = (\tilde{G}^{*L}, \tilde{G}^{*U})$ of IT2 FLS

$$D^{(2)}(C;A) = \frac{1}{\int_0^{h_A} f(\alpha)\,d\alpha} \int_0^{h_A} \sum_{i=1}^{n_\alpha} d_\theta^2(C, A_i^\alpha)\, p_i(\alpha)\, f(\alpha)\, d\alpha$$
$$= \frac{1}{\int_0^{h_A} f(\alpha)\,d\alpha} \int_0^{h_A} \sum_{i=1}^{n_\alpha} \left[\left(mid(C) - mid(A_i^\alpha)\right)^2 + \theta(\alpha)\left(spr(C) - spr(A_i^\alpha)\right)^2 \right] p_i(\alpha)\, f(\alpha) d\alpha$$

among all the intervals $C = [c_L, c_R]$, where, for each level α, the weights $p(\alpha) = (p_i(\alpha))_{i=1,\dots,n_\alpha}$ satisfy the properties $p_i(\alpha) \geq 0$ and $\sum_{i=1}^{n_\alpha} p_i(\alpha) = 1$, the weight function $f : [0,1] \to [0,+\infty[$ is such that $\int_0^{h_A} f(\alpha)\,d\alpha > 0$ and $\theta : [0,1] \to]0,1]$ is a function that indicates the relative importance of the spreads against the mids ([10,16]).

We have denoted by $mid(I) = (a+b)/2$ and $spr(I) = (b-a)/2$ the middle point and the spread of the interval $I = [a,b]$.

Theorem 1. *[4] The approximation interval $C^*(A) = C^*(A; p, f, \theta) = [c_L^*, c_R^*]$ of the T1 FQ A with respect to p, f, θ is given by*

$$c_L^* = \frac{\int_0^{h_A} \sum_{i=1}^{n_\alpha} mid(A_i^\alpha)\, p_i(\alpha)\, f(\alpha)\, d\alpha}{\int_0^{h_A} f(\alpha)\, d\alpha} - \frac{\int_0^{h_A} \sum_{i=1}^{n_\alpha} spr(A_i^\alpha)\, p_i(\alpha)\, f(\alpha)\, \theta(\alpha)\, d\alpha}{\int_0^{h_A} f(\alpha)\, \theta(\alpha)\, d\alpha}$$
$$c_R^* = \frac{\int_0^{h_A} \sum_{i=1}^{n_\alpha} mid(A_i^\alpha)\, p_i(\alpha)\, f(\alpha)\, d\alpha}{\int_0^{h_A} f(\alpha)\, d\alpha} + \frac{\int_0^{h_A} \sum_{i=1}^{n_\alpha} spr(A_i^\alpha)\, p_i(\alpha)\, f(\alpha)\, \theta(\alpha)\, d\alpha}{\int_0^{h_A} f(\alpha)\, \theta(\alpha)\, d\alpha}.$$

Definition 6. *We call evaluation of the T1 FQ A with respect to p, f, θ and $\lambda \in [0,1]$ the real number*

$$V^{\lambda,\theta}(A) = \phi_\lambda(C^*(A)),$$

where ϕ_λ is defined by $\phi_\lambda(I) = (1-\lambda)a + \lambda b = mid(I) + (2\lambda - 1)spr(I)$ for any interval $I = [a,b]$ and $\lambda \in [0,1]$ is a pessimistic/optimistic parameter. Thus

$$V^{\lambda,\theta}(A) = \frac{\int_0^{h_A} \sum_{i=1}^{n_\alpha} mid(A_i^\alpha)\, p_i(\alpha)\, f(\alpha)\, d\alpha}{\int_0^{h_A} f(\alpha)\, d\alpha} + (2\lambda - 1)\frac{\int_0^{h_A} \sum_{i=1}^{n_\alpha} spr(A_i^\alpha)\, p_i(\alpha)\, f(\alpha)\, \theta(\alpha)\, d\alpha}{\int_0^{h_A} f(\alpha)\, \theta(\alpha)\, d\alpha}.$$

This general formula includes, for suitable choices of parameters λ, p and f, the evaluations proposed by Fortemps and Roubens [8,3], Yager and Filev [18,19], Anzilli and Facchinetti [3] and Center of Gravity (COG), as shown in Table 1.

Definition 7. *We define the value of the IT2 FQ $\tilde{A} = (A^L, A^U)$ as*

$$V^{\lambda,\theta}(\tilde{A}) = (V^{\lambda,\theta}(A^L) + V^{\lambda,\theta}(A^U))/2.$$

As an application, we now compute the evaluation of the final output $\tilde{G}^* = (\tilde{G}^{*L}, \tilde{G}^{*U})$ given in (2) (see Fig. 4) using different methods. First, we evaluate the T1 FQs $\tilde{G}^{*L}$ and $\tilde{G}^{*U}$ and then we obtain the value of the IT2 FQ $\tilde{G}^*$ as $V(\tilde{G}^*) = (V(G^{*L}) + V(G^{*U}))/2$. The numerical results are shown in Table 2. The "Interval Type-2 Fuzzy Logic Toolbox" [6] produces 52 as centroid.

Table 1. Set of parameters

Evaluation	λ	$p_i(\alpha)$	$f(\alpha)$
Fortemps and Roubens	1/2	$1/n_\alpha$	n_α
Yager and Filev	1/2	$spr(A_i^\alpha)/\sum_{j=1}^{n_\alpha} spr(A_j^\alpha)$	1
Anzilli and Facchinetti	1/2	$spr(A_i^\alpha)/\sum_{j=1}^{n_\alpha} spr(A_j^\alpha)$	n_α
COG	1/2	$spr(A_i^\alpha)/\sum_{j=1}^{n_\alpha} spr(A_j^\alpha)$	$2\sum_{j=1}^{n_\alpha} spr(A_j^\alpha)$

Table 2. Evaluation of IT2 FQ $\tilde{G}^* = (\tilde{G}^{*L}, \tilde{G}^{*U})$

Evaluation	$V(G^{*L})$	$V(G^{*U})$	$V(\tilde{G}^*)$
Fortemps and Roubens	56.48	53.57	55.03
Yager and Filev	58.28	54.48	56.38
Anzilli and Facchinetti	56.49	53.53	55.01
COG	53.44	51.49	52.47

6 Conclusion

In this paper we introduce a different type-reduction method for IT2 FLSs. We consider only the T1 membership functions that bound the Output FOU zone and for its defuzzification we present a general formula completely different from centroid proposed by Karnik and Mendel for two reasons. First of all it is obtained working on an α-cuts approach while centroid works on x-axis. Moreover it is presented in a parametric formulation leaving a wide set of freedom. This opportunity has allowed us to obtain not only other methods already known, not only other completely new but the centroid too. We have obtained this general formula starting from an idea of the interval nearest to T1 FQ respect to a general functional suggested by the distance proposed by Trutschnig et al. [16]. Now we are working on a more general way to approximate T1 FQs based on a triangular fuzzy set and in the following on trapezoidal fuzzy sets. These works are in preparation.

References

1. Anzilli, L., Facchinetti, G.: Ambiguity of Fuzzy Quantities and a New Proposal for their Ranking. Przeglad Elektrotechniczny-Electrical Review 10, 280–283 (2012)
2. Anzilli, L., Facchinetti, G.: The total variation of bounded variation functions to evaluate and rank fuzzy quantities. International Journal of Intelligent Systems 28, 927–956 (2013)
3. Anzilli, L., Facchinetti, G., Mastroleo, G.: Evaluation and interval approximation of fuzzy quantities. In: Proceedings of 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2013), pp. 180–186. Atlantis Press (2013)
4. Anzilli, L., Facchinetti, G., Mastroleo, G.: A parametric approach to evaluate fuzzy quantities. Fuzzy Sets and Systems (2014), `http://dx.doi.org/10.1016/j.fss.2014.02.018`

5. Bojadziev, G., Bojadziev, M.: Fuzzy Logic for Business, Finance and Management, 2nd edn. World Scientific (2007)
6. Castillo, O.: Type-2 Fuzzy Logic in Intelligent Control Applications. STUDFUZZ, vol. 272. Springer, Heidelberg (2012)
7. Facchinetti, G., Pacchiarotti, N.: Evaluations of fuzzy quantities. Fuzzy Sets and Systems 157, 892–903 (2006)
8. Fortemps, P., Roubens, M.: Ranking and defuzzification methods based on area compensation. Fuzzy Sets and Systems 82, 319–330 (1996)
9. Grzegorzewski, P.: Nearest interval approximation of a fuzzy number. Fuzzy Sets and Systems 130, 321–330 (2002)
10. Grzegorzewski, P.: On the interval approximation of fuzzy numbers. In: Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R.R. (eds.) IPMU 2012, Part III. CCIS, vol. 299, pp. 59–68. Springer, Heidelberg (2012)
11. Karnik, N.N., Mendel, J.M.: Centroid of a type-2 fuzzy set. Inform. Sci. 132, 195–220 (2001)
12. Mendel, J.M.: Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice-Hall, Upper-Saddle River (2001)
13. Mendel, J.M.: Advances in type-2 fuzzy sets and systems. Information Sciences 177, 84–110 (2007)
14. Mendel, J.M., John, R.I.: Type-2 fuzzy sets made simple. IEEE Trans. Fuzzy Syst. 10(2), 117–127 (2002)
15. Mendel, J.M., John, R.I., Liu, F.L.: Interval type-2 fuzzy logical systems made simple. IEEE Transactions on Fuzzy Systems 14, 808–821 (2006)
16. Trutschnig, W., González-Rodrýguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences 179, 3964–3972 (2009)
17. Wu, D., Mendel, J.M.: Uncertainty measures for interval type-2 fuzzy sets. Information Sciences 177, 5378–5393 (2007)
18. Yager, R.R.: A procedure for ordering fuzzy subsets of the unit interval. Information Sciences 24, 143–161 (1981)
19. Yager, R.R., Filev, D.: On ranking fuzzy numbers using valuations. International Journal of Intelligent Systems 14, 1249–1268 (1999)
20. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning - 1. Information Sciences 8, 199–249 (1975)

Connecting Interval-Valued Fuzzy Sets with Imprecise Probabilities

Ignacio Montes, Enrique Miranda, and Susana Montes

University of Oviedo, Dept. of Statistics and O.R.,
C-Calvo Sotelo, s/n, 33007 Oviedo, Spain
{imontes,mirandaenrique,montes}@uniovi.es
http://www.unimode.es

Abstract. We study interval-valued fuzzy sets as a model for the imprecise knowledge of the membership function of a fuzzy set. We compare three models for the probabilistic information about this membership function: the set of distributions of the measurable selections, the upper and lower probabilities of the associated random interval, and its p-box. We give sufficient conditions for the equality between these sets, and establish a connection with the notion of probability induced by an intuitionistic fuzzy set. An alternative approach to the problem by means of sets of finitely additive distributions is also considered.

Keywords: Interval-valued fuzzy sets, random intervals, measurable selections, upper and lower probabilities, p-boxes.

1 Introduction

Interval-valued fuzzy sets [18] (IVF-sets, for short) were introduced as an extension of fuzzy sets [16] to model situations in which the "true" membership function is in some sense unknown. Then, instead of providing a precise membership degree, IVF-sets assign an interval of possible membership degrees. Thus, given an universe Ω, an IVF-set A is defined, for any $\omega \in \Omega$, by $[l_A(\omega), u_A(\omega)]$, and it is given the epistemic interpretation that all we know about the "true" membership degree of ω is that it belongs to that interval.

A related extension of fuzzy sets is given by *intuitionistic fuzzy sets* [1,2] (IF-sets, for short). For them, the interpretation is slightly different: they assign a membership and a non-membership degree to any element of the possibility space. Thus, an IF-set A is defined by two functions $\mu_{\mathrm{A}}, \nu_{\mathrm{A}} : \Omega \to [0,1]$, so that for any $\omega \in \Omega$, $\mu_{\mathrm{A}}(\omega)$ and $\nu_{\mathrm{A}}(\omega)$ model the degree in which ω satisfies and does not satisfy the notion encompassed by the fuzzy set A, with the restriction $\mu_{\mathrm{A}}(\omega) + \nu_{\mathrm{A}}(\omega) \leq 1$. In this sense, they constitute an instance of *bipolar* models [8]. Although IF-sets and IVF-sets model different situations, they are mathematically equivalent [4].

In this work, we shall assume that the IVF-set is defined on a probability space and that the unknown membership function is measurable, and shall investigate the probabilistic information about its associated distribution. In Section 3, we

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_6

compare three possible models, from the most to the least informative: the set of distributions of the measurable selections, those bounded between the upper and lower probabilities of the IVF-set, and those determined by the associated p-box. The advantage of these less informative models is that they are determined by a set and a point function, respectively.

We shall establish sufficient conditions for the equality between these three models, and give examples showing that they are not equivalent in general. Our results shall provide moreover a connection with the notion of probability induced by an intuitionistic fuzzy set from [10]. Finally, in Section 4 we shall investigate the problem without the assumption of measurability. In that case, we shall work with sets of finitely additive probabilities and with the theory of coherent lower previsions of Walley [15]. We shall see that the equivalences above do not always hold in this case. We conclude the paper with some additional remarks in Section 5. Due to the space restrictions, proofs have been omitted.

2 Preliminary Concepts

2.1 Random Sets

Since in this paper we shall deal with the probabilistic information of IVF-sets, it is interesting to recall a few notions of the sets of probabilities associated with multi-valued mappings. Given a probabilistic space $(\Omega, \mathcal{A}, P)$ and a measurable space $(\Omega', \mathcal{A}')$, a *random set* [6] is a multi-valued mapping $\Gamma : \Omega \to \mathcal{P}(\Omega')$ such that $\Gamma^*(A) = \{\omega \in \Omega : \Gamma(\omega) \cap A \neq \emptyset\} \in \mathcal{A}$ for any $A \in \mathcal{A}'$.

A random set Γ can be used to model the imprecise knowledge about a random variable X, in the sense that for every $\omega \in \Omega$ all we know about $X(\omega)$ is that it belongs to $\Gamma(\omega)$. Then, X belongs to the set of *measurable selections* of Γ:

$$S(\Gamma) = \{U : \Omega \to \Omega' \text{ random variable} \mid U(\omega) \in \Gamma(\omega) \ \forall \omega \in \Omega\}, \tag{1}$$

and the probability measure it induces on $\mathcal{A}'$ belongs to

$$\mathcal{P}(\Gamma) = \{P_U : U \in S(\Gamma)\}. \tag{2}$$

Another way of summarizing the information given by a random set is by means of its associated upper and lower probabilities:

Definition 1 ([6]). *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\Omega', \mathcal{A}')$ a measurable space and $\Gamma : \Omega \to \mathcal{P}(\Omega')$ a random set. Then its* upper *and* lower probabilities *$P^*_\Gamma, P_{*\Gamma} : \mathcal{A}' \to [0, 1]$ are given by:*

$$P^*_\Gamma(A) = P(\{\omega : \Gamma(\omega) \cap A \neq \emptyset\}) \textit{ and } P_{*\Gamma}(A) = P(\{\omega : \Gamma(\omega) \subseteq A\}) \ \forall A \in \mathcal{A}'. \tag{3}$$

These upper and lower probabilities define a credal set $\mathcal{M}(P^*_\Gamma)$ by:

$$\mathcal{M}(P^*_\Gamma) = \{P \text{ probability} : P_{*\Gamma}(A) \leq P(A) \leq P^*_\Gamma(A) \ \forall A \in \mathcal{A}'\}.$$

It is easy to see that $\mathcal{P}(\Gamma) \subseteq \mathcal{M}(P^*_\Gamma)$, and that the two sets do not coincide in general. The equality between them was investigated in [11] for the particular case of random closed intervals we shall consider later on.

2.2 P-Boxes

In these notes, we shall also work with one particular imprecise probability model: p-boxes.

Definition 2 ([9]). *A* distribution function *on $\Omega = [0,1]$ is a monotone mapping $F : [0,1] \to [0,1]$ that is right-continuous and satisfies $F(1) = 1$. Given two monotone functions $\underline{F}, \overline{F} : [0,1] \to [0,1]$ satisfying $\underline{F}(1) = \overline{F}(1) = 1$ and $\underline{F} \leq \overline{F}$, its associated* probability box *(p-box, for short) $(\underline{F}, \overline{F})$ is the set of distribution functions bounded between $\underline{F}$ and $\overline{F}$.*

The assumption of right-continuity of distribution functions guarantees that they are in a one-to-one correspondence with σ-additive probability measures. More generally, a monotone and normalized function $F : [0,1] \to [0,1]$ represents the cumulative probabilities associated with an infinite number of different finitely additive probability measures. See [14] for a study of p-boxes from the point of view of finitely additive probability measures.

The *credal set* associated with the p-box $(\underline{F}, \overline{F})$ is given by

$$\mathcal{M}(\underline{F}, \overline{F}) := \{P \text{ probability } : \underline{F} \leq F_P \leq \overline{F}\},$$

where F_P denotes the distribution function of P.

3 Probabilistic Information of Interval-Valued Fuzzy Sets

In this section, we detail a number of ways in which IVF-sets can be related to imprecise probability models.

3.1 IVFS as Random Intervals

As we mentioned in the introduction, an interval-valued fuzzy set can be regarded as a model for the imprecise knowledge of the membership function of a fuzzy set. In this section, we shall assume that the IVF-set is defined on the probability space $([0,1], \beta_{[0,1]}, \lambda_{[0,1]})$, and that the multi-valued mapping $\Gamma_A : [0,1] \to \mathcal{P}([0,1])$, given by

$$\Gamma_A(\omega) := [l_A(\omega), u_A(\omega)] \tag{4}$$

is a random set. This means [11, Theorem 3.1] that the mappings $l_A, u_A : [0,1] \to [0,1]$ must be $\beta_{[0,1]} - \beta_{[0,1]}$-measurable.

If we assume that the 'true' membership function imprecisely specified by means of the IVF-set is $\beta_{[0,1]} - \beta_{[0,1]}$-measurable, then it must belong to the set $S(\Gamma_A)$ given by Eq. (1), and its associated probability measure will belong to the set $\mathcal{P}(\Gamma_A)$ given by Eq. (2). As we have seen in Section 2.1, $\mathcal{P}(\Gamma_A)$ is included in the set $\mathcal{M}(P^*_{\Gamma_A})$ of probability measures that are dominated by $P^*_{\Gamma_A}$, but the two sets do not coincide in general. The equality between these two sets for random closed intervals was studied in [11]. Using the results from that paper, it is easy to establish the following:

Proposition 1. *$\mathcal{M}(P^*_{\Gamma_{\mathrm{A}}}) = \mathcal{P}(\Gamma_{\mathrm{A}})$ under any of the following conditions:*

(C1) l_A, u_A are increasing.
(C2) $l_A(\omega) = 0$ for any $\omega \in [0,1]$.
(C3) l_A, u_A are strictly comonotone[1]*.*

This result is interesting because it allows us to summarize the available information about the distribution of the membership function (the set of probability measures $\mathcal{P}(\Gamma_A)$) by means of the set function $P^*_{\Gamma_A}$. The conditions above may be interpreted in the following way:

(C1) As ω increases in [0,1], the evidence in favor of ω satisfying A increases.
(C2) There is no evidence supporting that any element satisfies A.
(C3) The intervals associated with the elements are ordered. In particular, this holds when the length of the intervals is constant.

On the other hand, the equality $\mathcal{P}(\Gamma) = \mathcal{M}(P^*_\Gamma)$ does not necessarily hold for all random closed intervals Γ [11, Example 3.3]. It is easy to adapt this example to our context and deduce that there are IVF-sets where the information about the membership function is not completely determined by the upper probability $P^*_{\Gamma_{\mathrm{A}}}$: it would suffice to take $\Gamma_{\mathrm{A}} : [0,1] \to \mathcal{P}([0,1])$ given by

$$\Gamma_{\mathrm{A}}(\omega) = \left[0.5 - \frac{\omega}{2}, 0.5 + \frac{\omega}{2}\right] \text{ for every } \omega \in [0,1]. \tag{5}$$

3.2 P-Boxes Induced by an IVF-Set

Now we take one step forward and study under which conditions the upper and lower probabilities $P^*_{\Gamma_A}, P_{*\Gamma_A}$ of the random interval associated with the IVF-set A can be summarized by means of two point functions: its lower and upper distribution functions $\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}} : [0,1] \to [0,1]$, given by

$$\underline{F}_{\mathrm{A}}(x) := P_{*\Gamma_{\mathrm{A}}}([0,x]) = P_{u_A}([0,x]) \text{ and } \overline{F}_{\mathrm{A}}(x) := P^*_{\Gamma_{\mathrm{A}}}([0,x]) = P_{l_A}([0,x]). \tag{6}$$

We shall refer to $(\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}})$ as the *p-box associated with the IVF-set A*. The credal set associated with this p-box is given by:

$$\mathcal{M}(\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}}) := \{Q : \beta_{[0,1]} \to [0,1] : \underline{F}_{\mathrm{A}}(x) \leq F_{\mathrm{Q}}(x) \leq \overline{F}_{\mathrm{A}}(x)\ \forall x \in [0,1]\},$$

where F_{Q} is the distribution function associated with the probability measure Q. It is immediate to see that the set $\mathcal{M}(\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}})$ includes $\mathcal{M}(P^*_{\Gamma_{\mathrm{A}}})$. However, the two sets do not coincide in general, and as a consequence the use of the lower and upper distribution functions may produce a loss of information. This was shown in [5, Example 3.3] for arbitrary random sets. Next, we give an example with random closed intervals associated with an IVF-set:

[1] We say that two functions $A, B : [0,1] \to [0,1]$ are *strictly comonotone* when $(A(\omega) - A(\omega')) \geq 0 \Leftrightarrow (B(\omega) - B(\omega')) \geq 0$ for any $\omega, \omega' \in [0,1]$.

Example 1. Consider the random set of Eq. (5), and let $(\underline{F}_A, \overline{F}_A)$ be its associated p-box. Given the distribution function F defined by:

$$F(x) = \begin{cases} \overline{F}_{\mathrm{A}}(x) & \text{if } x \leq \frac{1}{4}, \\ \frac{1}{2} & \text{if } x \in \left(\frac{1}{4}, \frac{3}{4}\right], \\ \underline{F}_{\mathrm{A}}(x) & \text{if } x > \frac{3}{4}, \end{cases}$$

its associated probability, P_F, belongs to $\mathcal{M}(\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}})$. However, P_F does not belong to $\mathcal{M}(P^*_{\Gamma_{\mathrm{A}}})$, because

$$P_F\left(\left[\frac{1}{4}, \frac{3}{4}\right]\right) = 0 < P_{*\Gamma_{\mathrm{A}}}\left(\left[\frac{1}{4}, \frac{3}{4}\right]\right) = \frac{1}{2}. \qquad \blacklozenge$$

Our next result shows that the sufficient conditions we have established in Proposition 1 for the equality $\mathcal{M}(P^*_{\Gamma_A}) = \mathcal{P}(\Gamma_A)$ also guarantee the equality between $\mathcal{M}(P^*_{\Gamma_{\mathrm{A}}})$ and $\mathcal{M}(\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}})$; thus, in those cases the p-box associated with the IVF-set keeps all the information about the probability distribution of the membership function.

Proposition 2. *Let A be a IVF-set on $([0,1], \beta_{[0,1]}, \lambda_{[0,1]})$, and let Γ_{A} be its associated random interval, given by Eq. (4). Then $\mathcal{P}(\Gamma_{\mathrm{A}}) = \mathcal{M}(\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}}) = \mathcal{M}(P^*_{\Gamma_{\mathrm{A}}})$ under any of the following conditions:*

(C1) l_A, u_A are increasing.
(C2) $l_A(\omega) = 0$ for every $\omega \in [0,1]$.
(C3) l_A and u_A are strictly comonotone.

3.3 Probabilities Associated with IFS

Another connection between probability theory and intuitionistic fuzzy sets was established in [10] by means of the probabilities induced by IF-sets. Given a probability space $(\Omega, \mathcal{A}, P)$, the probability associated with an IF-set A is an element of the interval

$$\left[\int_\Omega \mu_{\mathrm{A}} \mathrm{d}P, \int_\Omega 1 - \nu_{\mathrm{A}} \mathrm{d}P\right]. \tag{7}$$

This definition generalizes an earlier one by Zadeh [17]. Using this notion, in [10] a link is established with probability theory by considering the appropriate operators in the spaces of real intervals and of intuitionistic fuzzy sets. Note that it is assumed that we have a structure of probability space on Ω and that the functions $\mu_{\mathrm{A}}, \nu_{\mathrm{A}}$ are measurable, as we have done in this paper. From [4], it is known that IF-sets and IVF-sets are mathematically equivalent. In fact, given an IF-set with membership and non-membership functions μ_{A} and ν_{A}, it defines an IVF-set by considering $[\mu_{\mathrm{A}}(\omega), 1 - \nu_{\mathrm{A}}(\omega)]$ for every $\omega \in \Omega$.

If we assume that $(\Omega, \mathcal{A}, P) = ([0,1], \beta_{[0,1]}, \lambda_{[0,1]})$ and consider the random interval associated with the intuitionistic fuzzy set A interpreted as an IVF-set, we can see that the interval in Eq. (7) corresponds simply to the set of expectations of the measurable selections of Γ_{A}: it follows from [12, Thm. 14] that the *Aumann integral* [3] of $(id \circ \Gamma_{\mathrm{A}})$ satisfies

$$\left[\inf(A)\int(\mathrm{id}\circ\Gamma_{\mathrm{A}})\mathrm{d}P, \sup(A)\int(\mathrm{id}\circ\Gamma_{\mathrm{A}})\mathrm{d}P\right] = \left[(C)\int \mathrm{id}\ \mathrm{d}P^*_{\Gamma_{\mathrm{A}}}, (C)\int \mathrm{id}\ \mathrm{d}P_{*\Gamma_{\mathrm{A}}}\right]$$

where (C) denotes the *Choquet* integral [7] with respect to the non-additive measures $P_{*\Gamma_{\mathrm{A}}}, P^*_{\Gamma_{\mathrm{A}}}$, respectively. Since on the other hand it is immediate to see that

$$\sup(A)\int(\mathrm{id}\circ\Gamma_{\mathrm{A}})\mathrm{d}P = \int(1-\nu_{\mathrm{A}})\mathrm{d}P \text{ and } \inf(A)\int(\mathrm{id}\circ\Gamma_{\mathrm{A}})\mathrm{d}P = \int \mu_{\mathrm{A}}\mathrm{d}P,$$

we deduce that the probabilistic information about the intuitionistic fuzzy set A can be determined in particular by the lower and upper probabilities of its associated random interval.

4 A Non-measurable Approach

The previous developments assume that the IVF-set is defined on the probability space $([0,1], \beta_{[0,1]}, \lambda_{[0,1]})$ and that the functions $l_A, u_A : [0,1] \to [0,1]$, as well as the 'true' membership function of the fuzzy set modeled by A are $\beta_{[0,1]} - \beta_{[0,1]}$ measurable. Although this is a standard assumption when considering the probabilities associated with fuzzy events, it is arguably done for mathematical convenience only. In this section, we present an alternative approach where we get rid of the measurability assumptions by using finitely additive probabilities.

Consider thus a IVF-set A on $[0,1]$. Given its bounds l_A, u_A, we can define the multi-valued mapping $\Gamma_{\mathrm{A}} : [0,1] \to [0,1]$ by $\Gamma_{\mathrm{A}}(\omega) = [l_A(\omega), u_A(\omega)]\ \forall\omega$. Note that we are not assuming anymore that this multi-valued mapping is a random set. Our information about the 'true' membership function would be given by the set of functions

$$\{U : [0,1] \to [0,1] : l_A(\omega) \leq U(\omega) \leq u_A(\omega)\}. \tag{8}$$

Now, if we do not assume the measurability of l_A, u_A and consider then the field $\mathcal{P}([0,1])$ of all events in the initial space, we may not be able to model our uncertainty by means of a σ-additive probability measure. However, we can do so by means of a finitely additive probability measure. Moreover, the notions of lower and upper probabilities can be generalised to that case [13]. If for instance we consider a finitely additive probability P on $\mathcal{P}([0,1])$, then reasoning as in Section 3.1 we obtain that $P_U(C) \in [P_{\Gamma_{*A}}(C), P^*_{\Gamma_{\mathrm{A}}}(C)]$ for all $C \subseteq [0,1]$, where $P^*_{\Gamma_{\mathrm{A}}}, P_{*\Gamma_A}$ are defined by Eq. (3).

Then the information about P_U is given by the set of finitely additive probabilities dominated by $P^*_{\Gamma_{\mathrm{A}}}$, that coincides with the finitely additive probabilities induced by the elements of the set of Eq. (8). Hence, and in contradistinction to Section 3.1, when we work with finitely additive probabilities we do not need to make the distinction between $\mathcal{P}(\Gamma_{\mathrm{A}})$ and $\mathcal{M}(P^*_{\Gamma_{\mathrm{A}}})$.

The associated p-box is given now by the set of finitely additive distribution functions (that is, monotone and normalized) that lie between $\underline{F}_{\mathrm{A}}$ and $\overline{F}_{\mathrm{A}}$, where again $\underline{F}_{\mathrm{A}}, \overline{F}_{\mathrm{A}}$ are given by Eq. (6). Its associated set of finitely additive probability measures $\mathcal{M}(\underline{F}_A, \overline{F}_A)$ is determined by its lower envelope $\underline{E}_{(\underline{F},\overline{F})}$, that can be determined in the following way ([14]): if we denote by $\mathcal{H}$ the field of subsets of $[0,1]$ generated by the sets $\{[0,x],(x,1] : x \in [0,1]\}$, then any set in $\mathcal{H}$ is of the form

$$B_1 := [0,x_1] \cup (x_2,x_3] \cup \ldots (x_{2n},x_{2n+1}] \text{ or } B_2 := (x_1,x_2] \cup \ldots (x_{2n},x_{2n+1}]$$

for some $n \in \mathbb{N}, x_1 < x_2 < \cdots < x_n \in [0,1]$. It holds that

$$\begin{aligned}\underline{E}_{\underline{F},\overline{F}}(B_1) &= \underline{F}_{\mathrm{A}}(x_1) + \textstyle\sum_{i=1}^{n} \max\{0, \underline{F}_{\mathrm{A}}(x_{2i+1}) - \overline{F}_{\mathrm{A}}(x_{2i})\},\\ \underline{E}_{\underline{F},\overline{F}}(B_2) &= \textstyle\sum_{i=1}^{n} \max\{0, \underline{F}_{\mathrm{A}}(x_{2i}) - \overline{F}_{\mathrm{A}}(x_{2i-1})\},\end{aligned}$$

and also $\underline{E}_{\underline{F},\overline{F}}(C) = \sup_{B \subseteq C, B \in \mathcal{H}} \underline{E}_{\underline{F},\overline{F}}(B)$ for any $C \subsetneq [0,1]$.

Next, we investigate the equality $\mathcal{M}(P^*_{\Gamma_A}) = \mathcal{M}(\underline{F}_A, \overline{F}_A)$ under the conditions (C1)–(C3) considered in Proposition 1. We begin by showing that the two sets may not coincide when condition (C1) is satisfied.

Example 2. Consider the random interval defined by:

$$\Gamma_{\mathrm{A}}(\omega) = \begin{cases} [\omega, 2\omega] & \text{if } \omega \in \left[0, \frac{1}{3}\right] \\ \left[\frac{1}{3}, \frac{2}{3}\right] & \text{if } \omega \in \left(\frac{1}{3}, \frac{2}{3}\right] \\ [2\omega - 1, \omega] & \text{if } \omega \in \left(\frac{2}{3}, 1\right] \end{cases}$$

where in the initial space $([0,1], \mathcal{P}([0,1]))$ we consider a finitely additive probability P that agrees with $\lambda_{[0,1]}$ on $\beta_{[0,1]}$. Then, $P_{*\Gamma_{\mathrm{A}}}\left(\left[\frac{1}{3}, \frac{2}{3}\right]\right) = \frac{1}{3}$. However, it holds that:

$$\underline{E}_{\underline{F}_{\mathrm{A}},\overline{F}_{\mathrm{A}}}\left(\left[\frac{1}{3}, \frac{2}{3}\right]\right) = \underline{E}_{\underline{F}_{\mathrm{A}},\overline{F}_{\mathrm{A}}}\left(\left(\frac{1}{3}, \frac{2}{3}\right]\right) = \underline{F}_{\mathrm{A}}\left(\frac{2}{3}\right) - \overline{F}_{\mathrm{A}}\left(\frac{1}{3}\right) = \frac{2}{3} - \frac{2}{3} = 0. \blacklozenge$$

With respect to condition (C2), we have proven the following:

Proposition 3. *Let A be an IVF-set on $([0,1], \mathcal{P}([0,1]), P)$ with $l_A = 0$, and let $P_{*\Gamma_A}, (\underline{F}_A, \overline{F}_A)$ be its associated lower probability and p-box. Then, $\underline{E}_{\underline{F}_{\mathrm{A}},\overline{F}_{\mathrm{A}}} = P_{*\Gamma_{\mathrm{A}}}$.*

Finally, the equality does not hold for condition (C3), as we show next:

Example 3. Consider the random interval Γ_{A} defined on $([0,1], \mathcal{P}([0,1]), P)$ by:

$$\Gamma_{\mathrm{A}}(\omega) = \begin{cases} \left[\frac{1}{2} - \omega, 1 - \omega\right] & \text{if } \omega \in \left[0, \frac{1}{4}\right] \\ \left[\frac{1}{4}, \frac{3}{4}\right] & \text{if } \omega \in \left(\frac{1}{4}, \frac{3}{4}\right] \\ \left[\omega - \frac{1}{2}, \omega\right] & \text{if } \omega \in \left(\frac{3}{4}, 1\right], \end{cases}$$

and where P is a finitely additive probability that agrees with $\lambda_{[0,1]}$ on $\beta_{[0,1]}$. Since $l_A(\omega) = u_A(\omega) - \frac{1}{2}$, we see that l_A and u_A are strictly comonotone. If we consider the set $\left[\frac{1}{4}, \frac{7}{8}\right]$, we observe that

$$\underline{E}_{\underline{F}_{\mathrm{A}},\overline{F}_{\mathrm{A}}}\left(\left[\frac{1}{4}, \frac{7}{8}\right]\right) = \underline{E}_{\underline{F}_{\mathrm{A}},\overline{F}_{\mathrm{A}}}\left(\left(\frac{1}{4}, \frac{7}{8}\right]\right) = \frac{1}{4} < \frac{3}{4} = P_{*\Gamma_{\mathrm{A}}}\left(\left[\frac{1}{4}, \frac{7}{8}\right]\right). \qquad \blacklozenge$$

5 Conclusions

Our results show that the probabilistic information that an IVF-set holds about the underlying membership function can be summarized under some conditions by means of its associated p-box, although not in all cases. However, the correspondence depends on the measurability assumption of this membership function, and does not hold when we work with finitely additive probabilities instead.

In the future, we intend to deepen in the study of the imprecise probability models associated with an IVF-set, and to generalize our results to other possibility spaces. It would also be interesting to explore the alternative approach where no probability structure is considered in the initial space, and our knowledge is given instead by the set of possibility measures associated with the selections.

References

1. Atanassov, K.: Intuitionistic fuzzy sets. In: Proceedings of VII ITKR, Sofia (1983)
2. Atanassov, K.: Intuitionistic fuzzy sets. Fuz. Sets Syst. 20, 87–96 (1986)
3. Aumann, R.: Integrals of set valued functions. J. Math. Anal. Appl. 12, 1–12 (1965)
4. Bustince, H., Burillo, P.: Vague sets are intuitionistic fuzzy sets. Fuz. Sets and Syst. 79, 403–405 (1996)
5. Couso, I., Sánchez, L., Gil, P.: Imprecise distribution function associated to a random set. Inf. Sci. 159, 109–123 (2004)
6. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. 38, 325–339 (1967)
7. Denneberg, D.: Non-Additive Measure and Integral. Kluwer Academic, Dordrecht (1994)
8. Dubois, D., Prade, H.: Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets. Fuz. Sets Syst. 192, 3–24 (2012)
9. Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D., Sentz, K.: Constructing probability boxes and Dempster-Shafer structures. Technical Report SAND2002–4015, Sandia National Laboratories (2003)
10. Grzegorzewski, P., Mrowka, E.: Probability of intuitionistic fuzzy events. In: Grzegorzewski, P., Hryniewicz, O., Gil, M.A. (eds.) Soft Methods in Probability, Statistics and Data Analysis, pp. 105–115. Physica-Verlag (2002)
11. Miranda, E., Couso, I., Gil, P.: Random intervals as a model for imprecise information. Fuz. Sets Syst. 154, 386–412 (2005)
12. Miranda, E., Couso, I., Gil, P.: Approximations of upper and lower probabilities by measurable selections. Inf. Sci. 180, 1407–1417 (2010)
13. Miranda, E., de Cooman, G., Couso, I.: Lower previsions induced by multi-valued mappings. J. Stat. Plann. Inf. 133, 173–197 (2005)
14. Troffaes, M., Destercke, S.: Probability boxes on totally preordered spaces for multivariate modelling. Int. J. App. Reas. 52, 767–791 (2011)
15. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
16. Zadeh, L.: Fuzzy sets. Inf. Cont. 8, 338–353 (1965)
17. Zadeh, L.: Probability measures of fuzzy events. J. Math. Anal. Appl. 23, 421–427 (1968)
18. Zadeh, L.: The concept of a linguistic variable and its application to approximate reasoning II. Inf. Sci. 8, 301–357 (1975)

On Incomplete Label Ranking with IF-sets

Paweł P. Ładyżyński[1] and Przemysław Grzegorzewski[2,3]

[1] Interdisciplinary PhD Studies at the Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warsaw, Poland
pawelladyz@wp.pl

[2] Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw, Poland
pgrzeg@ibspan.waw.pl

[3] Faculty of Mathematics and Information Science,
Warsaw University of Technology,
Koszykowa 75, 00-662 Warsaw, Poland

Abstract. Probabilistic models, like the Mallows model, are commonly used for label ranking. However, for incomplete preferences the existing methods are exhaustive in the learning step and therefore the applications of the Mallows model in practical label ranking problems or in recommender systems are limited. In this paper, we show how to improve the Mallows model using IF-sets so it may become more simple and more effective for analyzing vague preferences and creating recommendations.

Keywords: IF-sets, incomplete data, instance-based learning, label ranking, the Mallows model, recommender systems.

1 Introduction

Label ranking is an important task in many applications like information retrieval, rating products or recommender systems. It can be treated as a generalization of a classification problem, where, instead of a ranking of all labels, only a single label is requested as a prediction for given observation. Thus, in brief, the label ranking can be perceived as a problem of learning a mapping from instances to rankings over a finite set of predefined labels.

This problem can be solved in different ways. Existing methods often use binary classification algorithms so the ranking is obtained by pairwise comparisons (see [6]). Another approaches utilize probabilistic models defined on a class of all rankings. As prominent example one can mention the Mallows model [7].

In recommender systems, due to the large amount of rated items, we typically meet incomplete preferences for all users available in data bases. However, it is not obvious, how to cope with such incomplete or vague preferences. Therefore, we propose an algorithm that combines the Mallows model with IF-set theory to get an effective method of label ranking and create recommendations in the presence incomplete rankings.

The paper is organized as follows. In Sec. 2 we describe briefly the problem of instance based label ranking and the Mallows model. Next, in Sec. 3 we show how

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_7

to apply IF-sets for modeling incomplete preferences and then how to enrich the classical Mallows model. Finally, in Sec. 4, we propose a new efficient algorithm for instance based label ranking and present results of the experimental study comparing our algorithm with other approaches.

2 Label Ranking and the Mallows Model

2.1 Basic Notions

Let $\mathbb{X}$, called an instance space, denote a set of elements (users, patients etc.) characterized by several attributes. Suppose that instead of classifying instances into separate classes, we associate each instance $x \in \mathbb{X}$ with a total order of all class labels $\mathbb{Y} = \{y_1, \ldots, y_M\}$. Moreover, we say that $y_i \succ_x y_j$ indicates that y_i is preferred to y_j given the instance x.

A total order $\succ_x$ can be identified with a permutation π_x of the set $\{1, \ldots, M\}$, where π_x is defined such that $\pi_x(i)$ is the index j of the class label y_j put on the i-th position in the order. Hence, $\pi_x^{-1}(j) = i$ gives the position of the j-th label (see [2]). The class of permutations of $\{1, \ldots, M\}$ will be denoted by Ω.

We may assume that every instance is associated with a probability distribution over Ω, i.e. for each instance $x \in \mathbb{X}$ there exists a probability distribution $\mathbb{P}(\cdot|x)$ such that, for every $\pi \in \Omega$, $\mathbb{P}(\pi|x)$ is the probability that $\pi_x = \pi$.

The main goal in label ranking is to predict a ranking of labels $y_1, \ldots, y_M$ for a new instance x, given some instances with known rankings of labels as a learning set. In practical issues, especially in recommender systems where the amount of available products is large, preference on instances known from the learning set do not usually contain all labels, i.e our information is of the form $y_{\pi_x(1)} \succ_x \ldots \succ_x y_{\pi_x(k)}$, where $k < M$.

To evaluate the predictive performance of a label ranker a suitable loss function on Ω is needed, e.g. based on Kendall's tau (see [2]).

2.2 The Mallows Model

Going back to the above mentioned probability distribution $\mathbb{P}(\cdot|x)$, we need a probabilistic model suitable for our considerations. In [2] the Mallows model was used in the context of an instance-based approach to label ranking.

The Mallows model is a distance-based probability model defined by

$$\mathbb{P}(\pi|\theta, \pi_0) = \frac{exp(-\theta D(\pi, \pi_0))}{\phi(\theta)}, \qquad (1)$$

where the ranking $\pi_0 \in \Omega$ is the location parameter (center ranking), D is a distance measure on rankings, $\phi = \phi(\theta)$ is a constant normalization factor and θ stands for a spread parameter which determines how quickly the probability decreases with the increasing distance between π and π_0.

The label ranking problem is then solved by the maximum likelihood estimation connected with (1). In [2] parameters θ, π_0 are estimated using $\pi_1, \ldots, \pi_k$

rankings connected with k nearest neighbors of a new instance x in the training set. It works nicely when all rankings from the training set are complete. Unfortunately, such situation is unusual in the real world problems.

To handle incomplete rankings in the training data it was proposed in [2] to maximize the probability

$$\mathbb{P}(\pi|\theta, \pi_0) = \sum_{\pi^* \in E(\pi)} \mathbb{P}(\pi^*|\theta, \pi_0), \tag{2}$$

where $E(\pi)$ - set of linear extensions of π. However, calculations with (2) are rather exhaustive. Therefore, we suggest below another method based on IF-modeling of incomplete rankings proposed by Grzegorzewski (see [3,4]).

3 IF-sets and Incomplete Preferences

Let $\mathbb{U}$ denote a usual set, called the universe of discourse. An IF-set (Atanassov's intuitionistic fuzzy set, see [1]) is given by a set of ordered triples $\tilde{C} = \{(u, \mu_{\tilde{C}}(u), \nu_{\tilde{C}}(u)) : u \in \mathbb{U}\}$, where $\mu_{\tilde{C}}, \nu_{\tilde{C}} : \mathbb{U} \to [0,1]$ stand for the membership and nonmembership functions, respectively. It is assumed that $0 \leq \mu_{\tilde{C}}(u) + \nu_{\tilde{C}}(u) \leq 1$ for each $u \in \mathbb{U}$.

In [3,4,5] Grzegorzewski proposed how to model preference systems admitting ties and missing ranks. The key idea is to represent a preference system by an appropriate IF-set. Consider any finite set of labels $\mathbb{Y} = \{y_1, \ldots, y_M\}$. Given any instance $x \in \mathbb{X}$ let us define two functions $w_x, b_x : \mathbb{Y} \to \{0, 1, \ldots, M-1\}$ as follows: for each $y_i \in \mathbb{Y}$ let $w_x(y_i)$ denote a number of elements in $\mathbb{Y}$ surely worse than y_i, while $b_x(y_i)$ let denote a number of elements surely better than y_i, with respect to the preference related to instance x. Next let

$$\mu_{\tilde{x}}(y_i) = \frac{w_x(y_i)}{M-1}, \quad \nu_{\tilde{x}}(y_i) = \frac{b_x(y_i)}{M-1}. \tag{3}$$

denote a membership and nonmembership function, respectively, of the IF-set $\tilde{x} = \{(y_i, \mu_{\tilde{x}}(y_i), \nu_{\tilde{x}}(y_i)) : y_i \in \mathbb{Y}\}$ describing the preference system connected with instance x.

Having any two instances $x_1, x_2 \in \mathbb{X}$ we may compute a correlation between preference systems $\tilde{x}_1, \tilde{x}_1$ generated by these instances, using the generalized Kendall's tau, admitting incomplete preferences (see [4]):

$$\begin{aligned}\tilde{\tau} = \frac{1}{2M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} [&sgn(\mu_{\tilde{x}_1}(y_j) - \mu_{\tilde{x}_1}(y_i)) \cdot sgn(\mu_{\tilde{x}_2}(y_j) - \mu_{\tilde{x}_2}(y_i)) \\ &+ sgn(\nu_{\tilde{x}_1}(y_j) - \nu_{\tilde{x}_1}(y_i)) \cdot sgn(\nu_{\tilde{x}_2}(y_j) - \nu_{\tilde{x}_2}(y_i))].\end{aligned} \tag{4}$$

In Sec. 2.1 we have identified preferences with an adequate permutation π_x of labels $\mathbb{Y}$. For possibly incomplete preferences we get incomplete permutation $\tilde{\pi} = \tilde{\pi}_x$ which might be identified with the corresponding IF-set $\tilde{x}$. Thus for any two instances $x_1, x_2 \in \mathbb{X}$ we have $\tilde{\tau} = \tilde{\tau}(\tilde{x}_1, \tilde{x}_2) = \tilde{\tau}(\tilde{\pi}_1, \tilde{\pi}_2)$. Hence, using (4), we may consider the following measure

$$D_{\tilde{\tau}}(\tilde{\pi}_1, \tilde{\pi}_2) = \frac{1 - \tilde{\tau}(\tilde{\pi}_1, \tilde{\pi}_2)}{2}, \tag{5}$$

which seems to be useful in the generalized Mallows model (1) admitting incomplete rankings and defined as follows

$$\tilde{\mathbb{P}}(\tilde{\pi}|\theta, \tilde{\pi}_0) = \frac{exp(-\theta D_{\tilde{\tau}}(\tilde{\pi}, \tilde{\pi}_0))}{\phi(\theta)}. \tag{6}$$

Of course, when modeling preferences by IF-sets one can also consider other substitutes for the measure D in (1), including different distances, dissimilarity measures or divergences (see, e.g., [8]). However, we have chosen a measure based on the generalized Kendall's tau because it is common to use distances utilizing the classical Kendall's coefficient in the Mallows model (see, e.g., [2]).

In the examples below we compare the suggested methodology with the results obtained using the distance based on the classical Kendall's tau for all linear extensions of incomplete rankings.

Example 1. Consider $M = 6$ labels and the following two ranking: $\pi_0 : y_1 \succ y_3 \succ y_4 \succ y_2 \succ y_5 \succ y_6$ and $\pi : y_3 \succ y_1 \succ y_5 \succ y_4 \succ y_6$. It is seen at once that the first ranking is complete, while the second one is incomplete because of unknown location of label y_2.

To perform the classical Mallows model one may consider possible six different locations of y_2 with respect to other labels. Using notation introduced in Sec. 2.1 if we put, e.g. $\pi^{-1}(2) = k$, which means that label y_2 is located on the k-th position in the complete ranking ($k = 1, \ldots, 6$), then for all labels y_j such that $\pi^{-1}(j) \geq k$, their position in the new ranking shifts to the right, so we get $\pi^{-1}(j) := \pi^{-1}(j) + 1$. The probabilities calculated for the classical Mallows model according to formula (1) for all possible location of the unknown label y_2 are given in Table 1. In these calculations the classical Kendall's τ was applied in (5) and the spread parameter $\theta = 1$ was assumed.

Table 1. Values of $\mathbb{P}(\pi|\theta, \pi_0)$ for different locations of y_2

$\pi^{-1}(2)$	1	2	3	4	5	6
$\mathbb{P}(\pi\|\theta, \pi_0)$	0.09049159	0.09673	0.1033985	0.09673	0.1033985	0.09673

On the other hand, we may construct IF-sets $\tilde{x}_0$ and $\tilde{x}$ describing preferences generated by π_0 and π, respectively. By (3) we get

$$\tilde{x}_0 = \{(y_1, 1, 0), (y_2, 0.4, 0.6), (y_3, 0.8, 0.2), (y_4, 0.6, 0.4), (y_5, 0.2, 0.8), (y_6, 0, 1)\}$$
$$\tilde{x} = \{(y_1, 0.6, 0.2), (y_2, 0, 0), (y_3, 0.8, 0), (y_4, 0.2, 0.6), (y_5, 0.4, 0.4), (y_6, 0, 0.8)\}$$

If we calculate probability (6) for the complete ranking π_0 and incomplete π using formula (5) based on the generalized Kendall's tau (4) then we get $\tilde{\mathbb{P}}(\pi|\theta, \pi_0) = 0.09673$. As we can see, (6) approximates possible probabilities quite well. We tried many other examples and the results were similarly good. □

Example 2. Now we will check what happen if there are more missing values in label ranking. Let us consider $M = 7$ labels and the following two ranking: $\pi_0 : y_1 \succ y_2 \succ y_3 \succ y_4 \succ y_5 \succ y_6 \succ y_7$ and $\pi : y_4 \succ y_1 \succ y_3 \succ y_7 \succ y_5$. So now the first ranking is complete, while the second one is incomplete because of two unknown location of labels y_2 and y_6.

Using the suggested methodology based on IF-sets and the generalized Kendall's tau (4) the probability value of (6) for the complete ranking π_0 and incomplete π equals $\tilde{\mathbb{P}}(\pi|\theta, \pi_0) = 0.05330688$.

However, if we apply the traditional approach based on possible linear extensions $\pi^* \in E(\pi)$ (see Sec. 2.2) we get $\min_{\pi^* \in E(\pi)}\{\mathbb{P}((\pi^*|\theta, \pi_0)\} = 0.0451233$ and $\max_{\pi^* \in E(\pi)}\{\mathbb{P}((\pi^*|\theta, \pi_0)\} = 0.0629746$, while the arithmetic mean and the median for all possible linear extensions $\{\mathbb{P}((\pi^*|\theta, \pi_0) : \pi^* \in E(\pi)\}$ equals 0.0551824 and 0.05459133, respectively. Hence again, IF-set based approach appears to be helpful in approximating the probability (1) for incomplete rankings. □

4 Incomplete Knowledge and the Mallows Model in Designing Recommendations

4.1 Main Idea

As we have mentioned above, our aim is to predict a ranking of labels for a given new instance x. Unfortunately, estimation of π from (6) is not very simple. However, in many applications it is not necessary to identify a whole ranking but it suffices to indicate only those labels which are located on the highest positions in the ranking. It is a typical case found in recommender systems.

In this contribution we apply the Mallows model to express the probability corresponding to the best label, i.e.

$$\tilde{\mathbb{P}}(y_j^{best}|\theta, \pi^*) = \frac{exp(-\theta D^*(y_j^{best}, y_j^{\pi^*}))}{\phi(\theta)}, \tag{7}$$

where D^* is the Euclidean distance between IF-sets given by

$$D^*(y_j^{best}, y_j^{\pi^*}) = \sqrt{\frac{1}{2}\sum_{i=1}^{n}((\mu_{y_j^{best}} - \mu_{\pi^*}(y_j))^2 + (\nu_{y_j^{best}} - \nu_{\pi^*}(y_j))^2)}. \tag{8}$$

In our case $\mu_{y_j^{best}} = 1$ and $\nu_{y_j^{best}} = 0$, as we want to calculate the probability that y_j is the best label for instance x. Then, as a final recommendation we assume

$$Y = \underset{y_j}{\operatorname{argmax}}\{\sum_{\pi^* \in \bar{\pi}_{kNN(x)}} \tilde{\mathbb{P}}(y_j^{best}|\theta, \pi^*)\}, \tag{9}$$

where $\bar{\pi}_{kNN(x)}$ is the set of preference systems connected with k instances nearest to x. To predict the complete ranking for instance x we order labels $y_1, \ldots, y_M$ according to the values of (7).

4.2 Algorithms

We propose two algorithms based on the ideas discussed above. The first one is a direct implementation of the method proposed in Sec. 4.1.

Mallows Best Probability Algorithm (MBP)

```
{Input: $x$ - new instance, $X$ - learning set of instances, $\bar{\pi}$ - labels connected with instances, $k$ - number of nearest neighbors}
1. Find k nearest neighbors of x in X.
2. For (j in 1 : M) calculate $\sum_{\pi^* \in \bar{\pi}_{kNN(x)}} \tilde{\mathbb{P}}(y_j^{best}|\theta, \pi^*)$
3. MBP-rank < − Sort labels according to the values obtained in step 2 (in case of ties a label with lower index is better in the ranking).
{Output: MBP-rank}
```

The second algorithm is a modification of MBP that replaces missing labels in $\bar{\pi}_{kNN(x)}$ by the most probable extension of $\pi^* \in \bar{\pi}_{kNN(x)}$ with respect to (1). This replacement idea was suggested in IBLR algorithm given in [2].

Multistep Mallows Best Probability Algorithm (MMBP)

```
{Input: $x$ - new instance, $X$ - learning set of instances, $\bar{\pi}$ - labels connected with instances, $k$ - number of nearest neighbors}
1. Find k nearest neighbors of x in X.
2. For (j in 1 : M) calculate $\sum_{\pi^* \in \bar{\pi}_{kNN(x)}} \tilde{\mathbb{P}}(y_j^{best}|\theta, \pi^*)$
3. MMBP-rank < − Sort labels according to the values obtained in step 2 (in case of ties a label with lower index is better in the ranking).
4. $\bar{\pi}^{mod}_{kNN(x)}$ < − Find the most probable extensions of $\pi^* \in \bar{\pi}_{kNN(x)}$ with respect to (6).
5. For (j in 1 : M) calculate $\sum_{\pi^*_{mod} \in \bar{\pi}^{mod}_{kNN(x)}} \tilde{\mathbb{P}}(y_j^{best}|\theta, \pi^*)$
6. MMBP-rankmod < − Sort labels according to the values obtained in step 5 (in case of ties a label with lower index is better in the ranking)
7. If (MMBP-rankmod ≠ MMBP-rank) then (MMBP-rank < − MMBP-rankmod, go to step 4) else (output(MMBP-rank)).
{Output: MMBP-rank}
```

4.3 Experimental Results

To evaluate the proposed method we compared it with the IBLR algorithm given in [2]. Two types of data sets were used in our experiment: (A) For classification data, we followed the procedure proposed in [2], i.e. the naive Bayes classifier was first trained on the complete data set and then, for each example, all the labels present in the data set were ordered with respect to the predicted class probabilities. (B) For regression data a certain number of (numerical) attributes was removed from the set of predictors and each one was considered as a label. To obtain a ranking, the attributes were standardized and then ordered (see [2]). To obtain incomplete ranks we changed some ranks in every ranking into NA (non available). We considered different proportions p of missing values.

To compare algorithms we used two quality measures: their prediction accuracy and the evaluation times. As a measure of accuracy we used Kendall's tau. We evaluated the experiments using leave-one-out crossvalidation and according to the random effect of removing labels from complete rankings we repeated the evaluation 20 times for every chosen value of p. The results shown in Table 2 and Table 3 are the mean results for a given p.

All evaluations were performed using R package. We set the number of nearest neighbors to 5 (function *knn* from FNN library). The evaluation times, i.e. times of one full leave-one-out crossvalidation procedure for every algorithm, are measured using *proc.time()*. In Table 2 and Table 3 we show the mean times for all evaluations. To improve performance and parallelize our calculations, we used library *snowfall* with parameters *sfInit(cpus=4, parallel=TRUE)* on Intel core i5 2450M CPU. All data sets used for experiments were downloaded from `http://www.uni-marburg.de/fb12/kebi/research/repository/`

Table 2. Comparison of label ranking algorithms for p = 30% missing labels in the learning set

	accuracy			time [s]		
data set	IBLR	MBP	MMBP	IBLR	MBP	MMBP
glass (A)	0.781	0.784	0.788	3.504	0.26	3.7
vowel (A)	0.817	0.795	0.819	102.03	1.05	102.26
housing (B)	0.670	0.665	0.670	8.44	0.70	8.95
elevators (B)	0.622	0.617	0.624	1371.86	225.83	1583.55
wisconsin (B)	0.432	0.420	0.427	316.12	0.40	319.54
average	0.664	0.656	0.665	360.39	45.65	403.60

Results given in Table 2 and Table 3 show that algorithms MBP, MMBP and IBLR have similar accuracy on our experimental sets. More precisely, MBP is usually slightly worse than the two other algorithms, but it is significantly faster. MMBP algorithm, which can be perceived as the improved (in some sense) MBP, behaves more or less like IBLR both with respect to the accuracy and evaluation time. Therefore, one may conclude that our IF-set based method for handling incomplete label ranking seems to be very promising: it might be as accurate as IBLR (in MMBP version), but if we allow a slight lower accuracy then, using MBP version, we get desired results much faster than using IBLR.

Table 3. Comparison of label ranking algorithms for $p = 50\%$ missing labels in the learning set

	accuracy			time [s]		
data set	IBLR	MBP	MMBP	IBLR	MBP	MMBP
glass (A)	0.688	0.685	0.687	5.12	0.29	5.42
vowel (A)	0.725	0.700	0.715	119.84	0.95	126.04
housing (B)	0.579	0.570	0.573	12.53	0.7	13.12
elevators (B)	0.540	0.530	0.535	2326.23	272.67	2598.56
wisconsin (B)	0.381	0.351	0.363	502.22	0.37	508.74
average	0.583	0.567	0.575	593.19	55.00	650.38

5 Conclusions

In practice, the choice of the best method should be determined by the data structure. In recommender systems the 2% better accuracy is not as crucial as the time performance. Moreover, obviously the time consumed by all this methods increases with the number of labels and the number of missing values. The typical situation in recommender systems is that the number of labeled products is very large and therefore most of labels are missing for each user. Thus, the proposed MBP algorithm seems to be a promising candidate for creating recommendations especially in the presence of large number of labeled items.

Acknowledgments. Study was supported by research fellowship within “Information technologies: research and their interdisciplinary applications” project co-financed by European Social Fund (agreement number POKL.04.01.01-00-051/10-00).

References

1. Atanassov, K.: Intuitionistic Fuzzy Sets: Theory and Applications. Springer-Verlag (1999)
2. Cheng, W., Dembczynski, K., Hüllermeier, E.: Decision Tree and Instance-Based Learning for Label Ranking. In: ICML 2009, Montreal (2009)
3. Grzegorzewski, P.: The coefficient of concordance for vague data. Computational Statistics & Data Analysis 51, 314–322 (2006)
4. Grzegorzewski, P.: Kendall's correlation coefficient for vague preferences. Soft Computing 13, 1055–1061 (2009)
5. Grzegorzewski, P., Ziembińska, P.: Spearman's rank correlation coefficient for vague preferences. In: Christiansen, H., De Tré, G., Yazici, A., Zadrozny, S., Andreasen, T., Larsen, H.L. (eds.) FQAS 2011. LNCS, vol. 7022, pp. 342–353. Springer, Heidelberg (2011)
6. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artificial Intelligence 172, 1897–1916 (2008)
7. Mallows, C.: Non-null ranking models. Biometrika 44, 114–130 (1957)
8. Montes, S., Iglesias, T.: Montes, S., Iglesias, T., Janiš, V., Montes, I.: A common framework for some comparison measures of IF-sets. In: IWIFSGN 2012, Warsaw (2012)

On the Continuity of Probability on *IF* Sets

Beloslav Riečan[1,2] and Alžbeta Michalíková[1]

[1] Faculty of Natural Sciences, Matej Bel University
Tajovského 40, Banská Bystrica, Slovakia
[2] Mathematical Institut, Slovak Academy of Sciences
Štefánikova 49, Bratislava, Slovakia
{Beloslav.Riecan,Alzbeta.Michalikova}@umb.sk

Abstract. Starting with a descriptive characterization of probability on the intuitionistic fuzzy sets, different formulations of continuity are presented. The main instrument is a Cignoli representation theorem on IF probabilities by classical Kolmogorovian probabilities.

Keywords: Probability, Fuzzy Sets, IF Sets.

1 Introduction

One of the most important results of mathematics in the 20th century is the Kolmogorov concept of probability based on the set theory (see e. g. [22] for a review). On the other hand a new point of view on the mathematical models of uncertainty has been given by the Zadeh fuzzy set theory ([23]). This theory generalizes the Borel classical set theory. And it is characteristic that one of the first basic result of the fuzzy school was devoted to some probabilistic aspects of the theory ([24]).

In the paper we are interested in the Atanassov intuitionistic fuzzy (IF) set theory ([1], [2]). Here the IF set is a pair $A = (\mu_A, \nu_A)$ of fuzzy sets such that $\mu_A + \nu_A \le 1$. And again, one of the first important direction of the IF set theory was in the studying of probability of IF sets. The probability on the family $\mathcal{F}$ of IF events has been defined in [11] as a mapping assigning to every IF set $A = (\mu_A, \nu_A)$ the interval

$$P(A) = \left[\int_\Omega \mu_A dP, \int_\Omega (1 - \nu_A) dP\right] \ .$$

Consequently probability has been defined axiomatically ([16])

$$P(A) = [P_1(A), P_2(A)] \ .$$

Let $\mathcal{I}$ be the family of all compact intervals in R. Since the mapping $P : \mathcal{F} \to \mathcal{I}$ is characterized by the functions $P_1, P_2 : \mathcal{F} \to [0,1]$, called states, the main results has been formulated for states. Recall that the additivity of the states has been taken with respect to the Lukasiewicz operations $\oplus, \odot$, and the continuity as the

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*, Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_8

upper continuity. In this paper some other forms of continuity are formulated and proved.

The main instrument of our investigations is the state representation theorem ([4], [5], [17], [19],[20]). By the theorem every IF state can be described by the help of the classical Kolmogorovian probabilities.

2 *IF* Spaces

Any subset A of a given space Ω can be identified with its characteristic function

$$I_A : \Omega \to \{0,1\}$$

where $I_A(\omega) = 1$, if $\omega \in A$, $I_A(\omega) = 0$, if $\omega \notin A$. From the mathematical point of view a fuzzy set is a natural generalization of I_A(see [23], [24]). It is a function

$$\varphi_A : \Omega \to [0,1] \ .$$

Evidently any set (i.e. two-valued function on $\Omega, I_A \to \{0,1\}$) is a special case of a fuzzy set (multi-valued function), $\varphi_A : \Omega \to [0,1]$.

There are many possibilities for characterizations of operations with sets (union $A \cup B$ and intersection $A \cap B$). We shall use so called Lukasiewicz characterization:

$$I_{A\cup B} = (I_A + I_B) \wedge 1,$$

$$I_{A\cap B} = (I_A + I_B - 1) \vee 0 \ .$$

(Here $(f \vee g)(\omega) = \max(f(\omega), g(\omega)), (f \wedge g)(\omega) = \min(f(\omega), g(\omega))$.) Hence if $\varphi_A, \varphi_B : \Omega \to [0,1]$ are fuzzy sets, then the union (disjunction φ_A or φ_B of corresponding assertions) can be defined by the formula

$$\varphi_A \oplus \varphi_B = (\varphi_A + \varphi_B - 1) \wedge 1,$$

the intersection (conjunction φ_A and φ_B of corresponding assertions) can be defined by the formula

$$\varphi_A \odot \varphi_B = (\varphi_A + \varphi_B - 1) \vee 0 \ .$$

In the paper we shall work with the Atanassov generalization of the notion of fuzzy set so-called IF-set (see [1], [2]), what is a pair

$$A = (\mu_A, \nu_A) : \Omega \to [0,1] \times [0,1]$$

of fuzzy sets $\mu_A, \nu_A : \Omega \to [0,1]$, where

$$\mu_A + \mu_A \leq 1 \ .$$

Evidently a fuzzy set $\varphi_A : \Omega \to [0,1]$ can be considered as an IF set, where $\mu_A = \varphi_A : \Omega \to [0,1], \nu_A = 1 - \varphi_A : \Omega \to [0,1]$. Here we have $\mu_A + \nu_A = 1$, while generally it can be $\mu_A(\omega) + \nu_A(\omega) < 1$ for some $\omega \in \Omega$. Geometrically an IF-set can be regarded as a function $A : \Omega \to \Delta$ to the triangle

$$\Delta = \{(u,v) \in R^2; 0 \leq u, 0 \leq v, u + v \leq 1\} \ .$$

Fuzzy set can be considered as a mapping $\varphi_A : \Omega \to D$ to the segment

$$D = \{(u,v) \in R^2; 0 \leq u \leq 1, 0 \leq v \leq 1, u + v = 1\}$$

and the classical set as a mapping $\psi : \Omega \to D_0$ from Ω to two-point set

$$D_0 = \{(0,1),(1,0)\} \ .$$

In the next definition we again use the Lukasiewicz operations.

Definition 1. *By an IF subset of a set Ω a pair $A = (\mu_A, \nu_A)$ of functions*

$$\mu_A : \Omega \to [0,1], \nu_A : \Omega \to [0,1]$$

is considered such that

$$\mu_A + \nu_A \leq 1 \ .$$

We call μ_A the membership function, ν_A the non membership function and

$$A \leq B \Longleftrightarrow \mu_A \leq \mu_B, \nu_A \geq \nu_B \ .$$

If $A = (\mu_A, \nu_A), B = (\mu_B, \nu_B)$ are two IF sets, then we define

$$\begin{aligned} A \oplus B &= ((\mu_A + \mu_B) \wedge 1, (\nu_A + \nu_B - 1) \vee 0), \\ A \odot B &= ((\mu_A + \mu_B - 1) \vee 0, (\nu_A + \nu_B) \wedge 1), \\ \neg A &= (1 - \mu_A, 1 - \nu_A) \ . \end{aligned}$$

Denote by $\mathcal{F}$ a family of IF sets such that

$$A, B \in \mathcal{F} \Longrightarrow A \oplus B \in \mathcal{F}, A \odot B \in \mathcal{F}, \neg A \in \mathcal{F} \ .$$

Example 1. Let $\mathcal{F}$ be the set of all fuzzy subsets of a set Ω. If $f : \Omega \to [0,1]$ then we define

$$A = (f, 1 - f),$$

i.e. $\nu_A = 1 - \mu_A$.

Example 2. Let $(\Omega, \mathcal{S})$ be a measurable space, $\mathcal{S}$ a σ-algebra, $\mathcal{F}$ the family of all pairs such that $\mu_A : \Omega \to [0,1], \nu_A : \Omega \to [0,1]$ are measurable. Then $\mathcal{F}$ is closed under the operations $\oplus, \odot, \neg$.

3 *IF* States

Definition 2. *Let $(\Omega, \mathcal{S})$ be a measurable space, hence Ω is a non-empty set, $\mathcal{S}$ is a σ-algebra of subsets of Ω. By $\mathcal{F}$ the family of all IF-sets $A = (\mu_A, \nu_A)$ is denoted with σ-measurable functions $\mu_A, \nu_A : \Omega \to [0,1]$.*

Definition 3. *Let $A = (\mu_A, \nu_A) \in \mathcal{F}, A_n = (\mu_{A_n}, \nu_{A_n}) \in \mathcal{F}, (n = 1, 2, ...)$. We write $A_n \nearrow A$ if $\mu_{A_n} \nearrow \mu_A, \nu_{A_n} \searrow \nu_A$. We write $A_n \searrow A$ if $\mu_{A_n} \searrow \mu_A, \nu_{A_n} \nearrow \nu_A$.*

Definition 4. *A mapping $m : \mathcal{F} \to [0,1]$ is called an IF state, if the following properties are satisfied*

(1.1) $m((0,1)) = 0, \quad m((1,0)) = 1,$
(1.2) $A \odot B = (0,1) \Longrightarrow m(A \oplus B) = m(A) + m(B),$
(1.3) $A_n \nearrow A \Longrightarrow m(A_n) \nearrow m(A)$.

Now the representation theorem of IF states will be presented ([4], [5], [18], [20]).

Theorem 1. *Let $m : \mathcal{F} \to [0,1]$ be an IF state. Then there exist probabilities $P, Q : \mathcal{S} \to [0,1]$ and $\alpha \in R$ such that*

$$m(A) = \int_\Omega \mu_A dP + \alpha \left(1 - \int_\Omega (\mu_A + \nu_A) dQ\right)$$

for all $A \in \mathcal{F}$.

Example 3. Let $\mathcal{F}$ be the set of all measurable fuzzy subsets in the measurable space $(\Omega, \mathcal{S}), m : \mathcal{F} \to [0,1]$ be a state, $A = (f, 1-f) \in \mathcal{F}$. Then

$$m(f) = \int_\Omega f dP + \alpha(1 - \int_\Omega (f + 1 - f) dQ) = \int_\Omega f dP \ .$$

4 Continuity

A simple consequence of additivity and upper continuity is lower continuity.

Theorem 2. *Let $A_n \in \mathcal{F}(n = 1, 2, ...), A_n \searrow A$. Then $m(A) = \lim_{n\to\infty} m(A_n)$.*

Proof. Put $B_n = \neg A_n (n = 1, 2, ...), B = \neg A$. Then

$$B \odot A = (1 - \mu_A \ , 1 - \nu_A) \odot (\mu_A, \nu_A) =$$
$$= ((1 - \mu_A + \mu_A - 1) \vee 0 \ , (1 - \nu_A + \nu_A) \wedge 1 = (0,1),$$
$$B \oplus A = ((1 - \mu_A + \mu_A) \wedge 1 \ , (1 - \nu_A + \nu_A - 1) \vee 0) = (1,0) \ .$$

Therefore

$$1 = m((1,0)) = m(B \oplus A) = m(B) + m(A),$$

hence

$$m(B) = 1 - m(A) \ .$$

Similarly

$$m(B_n) = 1 - m(A_n) \ .$$

Moreover

$$\mu_{B_n} = 1 - \mu_{A_n} \nearrow 1 - \mu_A = \mu_B,$$
$$\nu_{B_n} = 1 - \nu_{A_n} \searrow 1 - \nu_A = \nu_B \ .$$

Therefore $B_n \nearrow B$, and $m(B_n) \nearrow m(B)$, hence

$$m(A) = 1 - m(B) = 1 - \lim_{n\to\infty} m(B_n) =$$
$$= \lim_{n\to\infty} (1 - m(B_n)) = \lim_{n\to\infty} m(A_n) \ .$$

□

Of course, the continuity of IF states works not only in the monotone case.

Theorem 3. *Let* $A_n = (\mu_{A_n}, \nu_{A_n}) \in \mathcal{F}(n = 1, 2, ...), A = (\mu_A, \nu_A) \in \mathcal{F}$ *and*

$$\lim_{n\to\infty} \mu_{A_n} = \mu, \ \lim_{n\to\infty} (\mu_{A_n} + \nu_{A_n}) = \mu_A + \nu_A \ .$$

Then

$$\lim_{n\to\infty} m(A_n) = m(A) \ .$$

Proof. By Theorem 1 there exist probabilities $P, Q : \mathcal{S} \to [0, 1]$ and $\alpha \in R$ such that

$$m(A) = \int_\Omega \mu_A dP + \alpha \left(1 - \int_\Omega \mu_A + \nu_A dQ\right),$$
$$m(A_n) = \int_\Omega \mu_{A_n} dP + \alpha \left(1 - \int_\Omega \mu_{A_n} + \nu_{A_n} dQ\right) \ .$$

Of course, $(\mu_{A_n})_n, (\mu_{A_n} + \nu_{A_n})_n$ are bounded sequences of integrable functions and

$$\lim_{n\to\infty} \mu_{A_n} = \mu, \quad \lim_{n\to\infty} (\mu_{A_n} + \nu_{A_n}) = \mu_A + \nu_A \ .$$

Therefore by the Lebesgue integration theorem

$$\lim_{n\to\infty} m(A_n) = \lim \int_\Omega \mu_{A_n} dP + \alpha \left(1 - \lim_{n\to\infty} \int_\Omega (\mu_{A_n} + \nu_{A_n}) dQ\right) =$$
$$= \int_\Omega \mu_A dP + \alpha \left(1 - \int_\Omega (\mu_A + \nu_A) dQ\right) = m(A) \ .$$

□

The preceding theorem can be presented also in a more general form

Theorem 4. *Let $\mu_{A_n}, \nu_{A_n} \in \mathcal{F}, f, g : [0,1]^2 \to [0,1]$ be continuous functions, $f(u,v) + g(u,v) \leq 1$ for any $u, v \in [0,1]$. Let*

$$A_n = (f(\mu_{A_n}, \nu_{A_n}), g(\mu_{A_n}, \nu_{A_n}))\ (n = 1, 2, ...),\ A = (f(\mu_A, \nu_A), g(\mu_A, \nu_A))\ .$$

Then

$$\lim_{n \to \infty} m(A_n) = m(A)\ .$$

Proof. Evidently $A_n \in \mathcal{F}, A \in \mathcal{F}$. By Theorem 1

$$m(A_n) = \int_\Omega f(\mu_{A_n}, \nu_{A_n}) dP + \alpha(1 - \int_\Omega (f(\mu_{A_n}, \nu_{A_n}) + g(\mu_{A_n}, \nu_{A_n})) dQ),$$

hence by the Lebesgue integration theorem

$$\lim_{n \to \infty} m(A_n) =$$

$$= \lim_{n \to \infty} \int_\Omega f(\mu_{A_n}, \nu_{A_n}) dP + \alpha \left(1 - \lim_{n \to \infty} \int_\Omega (f(\mu_{A_n}, \nu_{A_n}) + g(\mu_{A_n}, \nu_{A_n})) dQ\right) =$$

$$= \int_\Omega f(\mu_A, \nu_A) dP + \alpha \left(1 - \int_\Omega (f(\mu_A, \nu_A) + g(\mu_A, \nu_A)) dQ\right) = m(A)\ .$$

□

Theorem 5. *Let $A_n \in \mathcal{F}$ $(n = 1, 2, \ldots), A_i \odot A_j = (0,1) (i \neq j)$. Put*

$$\bigoplus_{n=1}^{\infty} A_n = \bigvee_{n=1}^{\infty} \bigodot_{i=1}^{n} A_i\ .$$

Then

$$m\left(\bigoplus_{n=1}^{\infty} A_n\right) = \sum_{i=1}^{\infty} m(A_n)\ .$$

Proof. Put

$$S_n = \bigodot_{i=1}^{n} A_i\ \ (n = 1, 2, ...)\ .$$

Then $S_n \subset S_{n+1}, S_n \in \mathcal{F}$ $(n = 1, 2, ...), S_n \nearrow \bigoplus_{n=1}^{\infty} A_n$. Then

$$m\left(\bigoplus_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} m(S_n) = \lim_{n \to \infty} m\left(\bigodot_{i=1}^{n} A_i\right) =$$

$$= \lim_{n \to \infty} \sum_{i=1}^{n} m(A_i) = \sum_{i=1}^{\infty} m(A_n)\ .$$

□

5 Conclusions

The IF sets theory is important and useful so in some theoretical considerations (see e.g. [6], [13], [14]) as well as from the practial point of view (e.g. [2]). Moreover, all theorems stated above can be directly applied for the spaces of fuzzy sets.

On the other hand, the space of intuitionistic fuzzy sets can be embedded to a multivalued algebra, hence our results can be motivation for probability on MV-algebras ([21], [15]). Also some physical motivations and applications are possible (e. g. [8], [9]). On the base the Slovak school of D-posets ([7],[12]) as well as equivalent American theory of effect algebras ([10]) are available for further investigations.

Acknowledgments. The support of the grant VEGA 1/0120/14 is kindly announced.

References

1. Atanassov, K.T.: Intuitionistic Fuzzy Sets: Theory and Applications. STUDFUZZ. Physica Verlag, Heidelberg (1999)
2. Atanassov, K.T.: On Intuitionistic Fuzzy Sets. Springer, Berlin (2012)
3. Cignoli, L., D'Ottaviano, M., Mundici, D.: Algebraic Foundations of Many-valued Reasoning. Kluwer, Dordrecht (2000)
4. Ciungu, L., Riečan, B.: General form of probabilities on IF-sets. In: Fuzzy Logic and Applications. Proc. WILF Palermo, pp. 101–107 (2009)
5. Ciungu, L., Riečan, B.: Representation theorem for probabilities on IFS-events. Information Sciences 180, 793–798 (2010)
6. Ciungu, L., Kelemenová, J., Riečan, B.: A new point of view to the inclusion exclusion principle. In: 6th Int. Conf. on Intelligent Systems, IS 2012, Varna, Bulgaria, pp. 142–144 (2012)
7. Chovanec, F.: Difference Posets and their Graphical Representation. Liptovsk y Mikuláš (2014) (in Slovak)
8. Dvurečenskij, A., Pulmannová, S.: New Trends in Quantum Structures. Kluwer, Dordrecht (2000)
9. Dvurečenskij, A., Rachunek: Riečan and Bosbach states for bounded noncommutative RI-monoids. Math. Slovaca 56, 487–500 (2006)
10. Foulis, D., Bennett, M.: Efect algebras and unsharp quantum logics. Found. Phys. 24, 1325–1346 (1994)
11. Grzegorzewski, P., Mrowka, E.: Probability of intuistionistic fuzzy events. In: Grzegorzewski, P., et al. (eds.) Soft Metods in Probability, Statistics and Data Analysis, pp. 105–115 (2002)
12. Kopka, F., Chovanec, F.: D-posets. Math. Slovaca 44, 21–34 (1994)
13. Lendelová, K.: A note on invariant observables. International Journal of Theoretical Physics 45, 915–923 (2006)
14. Michalíková, A.: Absolute value and limit of the function defined on IF sets. Notes on Intuitionistic Fuzzy Sets 18, 8–15 (2012)
15. Montagna, F.: An algebraic approach to propositional fuzzy logic. J. Logic Lang. Inf (D. Mundici et al. eds.), Special issue on Logics of Uncertainty 9, 91–124 (2000)

16. Riečan, B.: A descriptive definition of the probability on intuitionistic fuzzy sets. In: Wagenecht, M., Hampet, R. (eds.) EUSFLAT 2003, pp. 263–266 (2003)
17. Riečan, B.: Representation of probabilities on IFS events. In: Lopez-Diaz, et al. (eds.) Soft Methodology and Random Information Systems, pp. 243–248 (2004)
18. Riečan, B.: On a problem of Radko Mesiar: general form of IF-probabilities. Fuzzy Sets and Systems 152, 1485–1490 (2006)
19. Riečan, B.: Probability theory on intuitionistic fuzzy events. In: Aguzzoli, S., et al. (eds.) Algebraic and Proof theoretic Aspects of Non-Classical Logic, Papers in Honour of Daniele Mundici's 60th Birthday. LNCS, pp. 290–308. Springer, Heidelberg (2007)
20. Riečan, B.: Analysis of Fuzzy Logic Models. In: Koleshko, V.M. (ed.) Intelligent Systems, pp. 219–244. INTECH (2012)
21. Riečan, B., Mundici, D.: Probability in MV-algebras. In: Pap, E. (ed.) Handbook of Measure Theory II, pp. 869–910. Elsevier, Heidelberg (2002)
22. Riečan, B., Neubrunn, T.: Integral, Measure and Ordering. Kluwer, Dordrecht (1997)
23. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–358 (1965)
24. Zadeh, L.A.: Probability measures on fuzzy sets. J. Math. Abal. Appl. 23, 421–427 (1968)

Bayesian Updating under Incomplete or Imprecise Information in Finite Spaces

Giulianella Coletti[1], Davide Petturiti[2], and Barbara Vantaggi[2]

[1] Dip. Matematica e Informatica, Università di Perugia, Italy
giulianella.coletti@unipg.it
[2] Dip. S.B.A.I., Università di Roma "La Sapienza", Italy
{barbara.vantaggi,davide.petturiti}@sbai.uniroma1.it

Abstract. We provide (in a finite setting) a closed form expression for the lower envelope of the set of all the possible Bayesian posteriors derivable from a possibly incomplete or imprecise prior distribution (giving rise to a 2-monotone capacity) and a likelihood function.

Keywords: Bayesian updating, coherence, conditional probability, belief function, 2-monotone capacity.

1 Introduction

The classical Bayesian paradigm relies on a precise and complete probabilistic *prior* and *likelihood* assessment $\{P(H_i), P(E|H_i)\}_{i=1,\dots,n}$ and gives rise to a unique *posterior* distribution $\{P(H_i|E)\}_{i=1,\dots,n}$, whenever $P(E) > 0$. However, in real applications (e.g., medical diagnosis, forensic analysis and legal processes, to cite some) the prior knowledge could be imprecise (e.g., a belief function) or, even if precise, it could be only partially specified or defined on different hypotheses. At the same time, the expert could be interested in Bayesian queries on events more complex than the $H_i|E$'s.

The cases described above induce a (convex) set of prior probabilities whose lower envelope turns out to be a *belief function* [12,20,14,6]. Hence, the problem of non-unicity of the posterior needs to be dealt referring to the entire class of probabilistic extensions, and a characterization of the envelopes of such set is desirable, especially with a *sensitivity analysis* in view.

The main aim of this paper is to prove a generalized version of Bayes' theorem for finite spaces when the prior information is expressed by a 2-monotone capacity on the algebra spanned by the H_i's and the statistical model is still a likelihood function on the events $E|H_i$'s. Actually, our results can be generalized (see [5]) in order to extend results proved in [25,26], by allowing conditioning to any event in the algebra $\mathcal{A}$ spanned by E and the H_i's, without any positivity assumption on the corresponding (lower or upper) probability. This aim is in line with that of Walley [24].

Our contribution consists in providing a closed form expression for the lower envelope of the set of full conditional probabilities on $\mathcal{A}$ extending a complete

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_9

and precise prior probability and a likelihood function. Then we characterize the lower envelope of the coherent conditional probability extensions of a prior probability referring to events different from those where the likelihood is given. Finally, a generalization of the first result is proved, by considering a prior 2-monotone capacity and a likelihood function. We show that the "lower posterior probability" may fail 2-monotonicity: in the case the lower posterior probability is a 2-monotone capacity, then the updating procedure can be iterated.

2 Framework of Reference

Let $\mathcal{A}$ be a Boolean algebra of *events*, endowed with the usual Boolean operations of contrary $(\cdot)^c$, disjunction $\vee$, and conjunction $\wedge$, and the partial order of implication $\subseteq$. We denote with Ω and $\emptyset$, respectively, the *sure event* and the *impossible event* which coincide with the top and the bottom elements of $\mathcal{A}$, respectively. A subset $\mathcal{H} \subseteq \mathcal{A}^0 = \mathcal{A} \setminus \{\emptyset\}$ is said an *additive class* if it is closed under finite disjunctions.

We refer to the following axiomatic definition of *conditional probability* [7] which is equivalent to [10,9].

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

(i) $P(E|H) = P(E \wedge H|H)$, for every $E \in \mathcal{A}$ and $H \in \mathcal{H}$;
(ii) $P(\cdot|H)$ is a finitely additive probability on $\mathcal{A}$, for any $H \in \mathcal{H}$;
(iii) $P(E \wedge F|H) = P(E|H) \cdot P(F|E \wedge H)$, for any $H, E \wedge H \in \mathcal{H}$ and $E, F \in \mathcal{A}$.

Following [13], we say that a conditional probability $P(\cdot|\cdot)$ is *full on $\mathcal{A}$* if $\mathcal{H} = \mathcal{A}^0$. In order to deal with an assessment P on an *arbitrary* set $\mathcal{G}$ of conditional events we need to resort to the concept of *coherence* [7] (equivalent to [27,17]).

Definition 2. *Given an arbitrary set $\mathcal{G} = \{E_i|H_i\}_{i \in I}$ of conditional events, an assessment $P : \mathcal{G} \to [0,1]$ is a* **coherent conditional probability** *if and only if there is a conditional probability $\tilde{P} : \mathcal{A} \times \mathcal{H} \to [0,1]$ with $\mathcal{A} \times \mathcal{H} \supseteq \mathcal{G}$ extending the assessment P (i.e., $\tilde{P}_{|\mathcal{G}} = P$).*

By the conditional version [27,17] of *de Finetti's fundamental theorem for probabilities* [11], any coherent conditional probability P on $\mathcal{G}$ can be extended coherently to any further set $\mathcal{G}' \supset \mathcal{G}$ of conditional events. In general, the extension on $\mathcal{G}'$ is not unique thus we consider the set $\mathcal{P} = \{\tilde{P}(\cdot|\cdot)\}$ of all the coherent extensions of P. Such set is a compact subset of the space $[0,1]^{\mathcal{G}'}$ endowed with the product topology of pointwise convergence and is the Cartesian product of (possibly degenerate) closed intervals, which determine the *lower and upper envelopes* $\underline{P} = \min \mathcal{P}$ and $\overline{P} = \max \mathcal{P}$, where the minimum and the maximum are intended pointwise on the elements of $\mathcal{G}'$. The functions $\underline{P}$ and $\overline{P}$ on $\mathcal{G}'$ are *coherent lower and upper conditional probabilities* [7], respectively.

Notice that $\underline{P}$ and $\overline{P}$ are *dual*, i.e., $\overline{P}(E|H) = 1 - \underline{P}(E^c|H)$ if $E|H, E^c|H \in \mathcal{G}'$, thus, when $\mathcal{G}'$ is a structured set $\mathcal{A} \times \mathcal{H}$ the knowledge of $\underline{P}$ (simply called *lower conditional probability* in this case) is sufficient to recover $\overline{P}$.

Recall that a lower conditional probability $\underline{P}$ on $\mathcal{A} \times \mathcal{H}$ is such that for every $H \in \mathcal{H}$, $\underline{P}(\emptyset|H) = 0$, $\underline{P}(\Omega|H) = 1$, $\underline{P}(E|H) = \underline{P}(E \wedge H|H)$ and $\underline{P}(\cdot|H)$ is superadditive on $\mathcal{A}$. Furthermore, for $H \in \mathcal{H}$, $\underline{P}(\cdot|H)$ is said n-monotone ($n \geq 2$) on $\mathcal{A}$ if

$$\underline{P}\left(\bigvee_{i=1}^{n} E_i \,\middle|\, H\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} \underline{P}\left(\bigwedge_{i \in I} E_i \,\middle|\, H\right), \tag{1}$$

for every $E_1, \ldots, E_n \in \mathcal{A}$. In particular, for $H \in \mathcal{H}$, $\underline{P}(\cdot|H)$ is said a *belief function* [20] on $\mathcal{A}$ if it is n-monotone for every $n \geq 2$.

3 Precise and Complete Prior and Likelihood Function

Let $\mathcal{L} = \{H_1, \ldots, H_n\}$ be a finite partition of Ω, E an arbitrary possible event, and $\mathcal{A} = \langle \{E\} \cup \mathcal{L} \rangle$ the algebra spanned by $\{E\} \cup \mathcal{L}$, whose set of atoms is $\mathcal{C}_{\mathcal{A}}$.

A *likelihood function* f (see, e.g., [4]) is *any* map from $\{E\} \times \mathcal{L}$ to $[0,1]$, with the only constraint that $f(E|H_i) = 0$ if $E \wedge H_i = \emptyset$ and $f(E|H_i) = 1$ if $E \wedge H_i = H_i$.

Given a likelihood function $f(E|\cdot)$ and a *prior probability distribution* $p(\cdot)$ on $\mathcal{L}$, the joint assessment $\{p, f\}$ is a coherent conditional probability on $\mathcal{G} = \{E|H_i, H_i\}_{i=1,\ldots,n}$ [18,7,22] which determines a unique coherent extension P on $\mathcal{G}' = \mathcal{A} \times (\{\Omega\} \cup \mathcal{L})$. Nevertheless, the further extension of P on $\mathcal{A} \times \mathcal{A}^0$ is not unique in general so we need to consider the set

$$\mathcal{P} = \{\tilde{P} \,:\, \text{full conditional probability on } \mathcal{A} \text{ s.t. } \tilde{P}_{|\mathcal{G}'} = P\}.$$

The following theorem provides a closed form expression for $\underline{P} = \min \mathcal{P}$.

Theorem 1. *Given a likelihood function $f(E|\cdot)$ and a prior probability distribution $p(\cdot)$ on $\mathcal{L}$, for every $F|K \in \mathcal{A} \times \mathcal{A}^0$, $\underline{P}(F|K) = 1$ when $F \wedge K = K$, and if $F \wedge K \neq K$, then:*

(i) if $P(K) > 0$ then

$$\underline{P}(F|K) = \frac{P(F \wedge K)}{P(K)}; \tag{2}$$

(ii) if $P(K) = 0$, then if $I \neq \emptyset$ and $H_j \wedge F \wedge K \neq \emptyset$ for all $j \in J$ and $F^c \wedge K \wedge \left(\bigvee_{i \in I} H_i\right)^c = \emptyset$, where $I, J \subseteq \{1, \ldots, n\}$ are, respectively, the maximum and minimum index set such that $\bigvee_{i \in I} H_i \subseteq K \subseteq \bigvee_{j \in J} H_j$, then

$$\underline{P}(F|K) = \min\left\{ \min_{\substack{E \wedge H_i \subseteq F \wedge K \\ E^c \wedge H_i \subseteq F^c \wedge K}} f(E|H_i), \min_{\substack{E^c \wedge H_i \subseteq F \wedge K \\ E \wedge H_i \subseteq F^c \wedge K}} (1 - f(E|H_i)) \right\}; \tag{3}$$

otherwise $\underline{P}(F|K) = 0$.

Proof. The proof is trivial in the case $F \wedge K = K$ or $P(K) > 0$. Assume $F \wedge K \neq K$ and $P(K) = 0$. Denote with $E^\bullet$ either E or E^c and let $\mathcal{C}_1 = \{E^\bullet \wedge H_i \in \mathcal{C}_\mathcal{A} : P(E^\bullet \wedge H_i) = 0\}$. The lower bound $\underline{P}(F|K)$ can be computed by solving the optimization problem (see [7,1]) with non-negative unknowns x_j^1 for $E \wedge H_j \in \mathcal{C}_1$, $j \in J$, and y_j^1 for $E^c \wedge H_j \in \mathcal{C}_1$, $j \in J$,

$$\text{minimize} \left[\sum_{E \wedge H_j \subseteq F \wedge K} x_j^1 + \sum_{E^c \wedge H_j \subseteq F \wedge K} y_j^1 \right]$$

$$\begin{cases} x_j^1 = f(E|H_j) \cdot (x_j^1 + y_j^1) & \text{if } E \wedge H_j \in \mathcal{C}_1 \text{ and } E^c \wedge H_j \in \mathcal{C}_1 \text{ and } j \in J, \\ \sum\limits_{E \wedge H_j \subseteq K} x_j^1 + \sum\limits_{E^c \wedge H_j \subseteq K} y_j^1 = 1. \end{cases}$$

The unknowns in the system are divided in independent groups corresponding to each H_j with $j \in J$ and are constrained together only by the last equation. If $I = \emptyset$ or there exits $j \in J$ s.t. $H_j \wedge F \wedge K = \emptyset$ or $F \wedge K \wedge \left(\bigvee_{i \in I} H_i\right)^c \neq \emptyset$, one can always build a solution such that $\sum\limits_{E \wedge H_j \subseteq F \wedge K} \mathbf{x}_j^1 + \sum\limits_{E^c \wedge H_j \subseteq F \wedge K} \mathbf{y}_j^1 = 0$ and $\sum\limits_{E \wedge H_j \subseteq F^c \wedge K} \mathbf{x}_j^1 + \sum\limits_{E^c \wedge H_j \subseteq F^c \wedge K} \mathbf{y}_j^1 = 1$, which implies $\underline{P}(F|K) = 0$. In the opposite case the minimum is achieved in correspondence of those solutions such that $\mathbf{x}_i^1 + \mathbf{y}_i^1 = 1$ for $E^\bullet \wedge H_i \subseteq F \wedge K$ and $(E^\bullet)^c \wedge H_i \subseteq F^c \wedge K$, thus the conclusion follows. □

Let us note that if $P(K) > 0$, $\underline{P}(\cdot|K)$ is a probability measure (and so a belief function) on $\mathcal{A}$. However the following example shows that for some $K \in \mathcal{A}^0$ with $P(K) = 0$, the lower envelope $\underline{P}(\cdot|K)$ can fail even 2-monotonicity.

Example 1. Let $\mathcal{L} = \{H_1, H_2, H_3, H_4\}$ be a partition of Ω and E an event logically independent of $\mathcal{L}$. Consider the likelihood $f(E|H_i) = \frac{1}{2}$, $i = 1, 2, 3, 4$, and the prior probability distribution $p(H_1) = 1$ and $p(H_i) = 0$, $i = 2, 3, 4$.

Let $K = H_2 \vee H_3 \vee H_4$ and $F = (E \wedge H_2) \vee (E^c \wedge H_3) \vee (E \wedge H_4)$. It holds $\underline{P}(E \vee F|K) = \underline{P}(E|K) = \underline{P}(F|K) = \frac{1}{2}$ and $\underline{P}(E \wedge F|K) = 0$, which implies $\underline{P}(\cdot|K)$ is not 2-monotone on $\mathcal{A} = \langle \{E\} \cup \mathcal{L} \rangle$ since it is $\underline{P}(E \vee F|K) < \underline{P}(E|K) + \underline{P}(F|K) - \underline{P}(E \wedge F|K)$.

4 Imprecise or Partial Prior Information

Consider two finite Boolean algebras of events $\mathcal{A}, \mathcal{A}'$, and a probability measure P on $\mathcal{A}$. If the algebra of interest is $\mathcal{A}'$ we can consider the set of coherent extensions on $\mathcal{G}' = (\mathcal{A} \times \{\Omega\}) \cup (\mathcal{A}' \times \mathcal{A}'^0)$

$$\mathcal{P} = \{\tilde{P} : \text{coherent conditional probability on } \mathcal{G}' \text{ s.t. } \tilde{P}_{|\mathcal{A} \times \{\Omega\}} = P\}$$

with its lower envelope $\underline{P} = \min \mathcal{P}$. Next theorem provides a closed form expression for $\underline{P}$ on $\mathcal{A}' \times \mathcal{A}'^0$, relying on the lower and upper probabilities $\underline{P}(\cdot) = \underline{P}(\cdot|\Omega)$ and $\overline{P}(\cdot) = \overline{P}(\cdot|\Omega)$ on $\mathcal{A}'$, obtained extending P on $\mathcal{A} \cup \mathcal{A}'$, which are known to be, respectively, a belief function and a plausibility function [14].

Theorem 2. *Let $\mathcal{A}, \mathcal{A}'$ be two finite Boolean algebras, P a probability measure on $\mathcal{A}$, and $\underline{P}(\cdot|\cdot)$ the lower envelope of the set of coherent extensions of P on $\mathcal{G}'$. The following statements hold:*

(i) *$\underline{P}(\cdot|K)$ is a belief function on $\mathcal{A}'$, for every $K \in \mathcal{A}'^0$;*
(ii) *for every $F|K \in \mathcal{A}' \times \mathcal{A}'^0$, $\underline{P}(F|K) = 1$ when $F \wedge K = K$, and if $F \wedge K \neq K$, then we have*

$$\underline{P}(F|K) = \begin{cases} \frac{\underline{P}(F\wedge K)}{\underline{P}(F\wedge K)+\overline{P}(F^c\wedge K)} & \text{if } \underline{P}(F \wedge K) + \overline{P}(F^c \wedge K) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Proof. We prove condition *(ii)* first. If $F \wedge K = K$, for every $\tilde{P} \in \mathcal{P}$, $\tilde{P}(F|K) = 1$, so $\underline{P}(F|K) = 1$. Hence assume $F \wedge K \neq K$. By Proposition 3.1 in [14], $\underline{P}(\cdot)$ is a belief function on $\mathcal{A}'$, so Theorem 7.2 in [23] implies equation (4) when $\underline{P}(F \wedge K) + \overline{P}(F^c \wedge K) > 0$. Finally, in the case $\underline{P}(F \wedge K) + \overline{P}(F^c \wedge K) = 0$ equation (4) follows by Proposition 3 in [1].

Now we prove condition *(i)*. Theorem 1 in [15] (or, equivalently, Theorem 4.1 in [21]) implies that $\underline{P}(\cdot|K)$ is a belief function on $\mathcal{A}'$ when $\underline{P}(K) > 0$, which implies $\underline{P}(F \wedge K) + \overline{P}(F^c \wedge K) > 0$. When $\underline{P}(K) = 0$, the claim follows by the monotonicity of $\underline{P}(\cdot|K)$ and since $\underline{P}(F|K) > 0$ only for events $F \in \mathcal{A}'$ such that $F \wedge K = K$. □

Previous theorem differs from Theorem 7.2 in [23], where $\underline{P}(F|K)$ is not defined when $\overline{P}(K) = 0$, moreover, in the case $\overline{P}(K) > 0$ and $\underline{P}(F \wedge K) + \overline{P}(F^c \wedge K) = 0$, $\underline{P}(F|K)$ is set equal to 1, which is not the minimum coherent value for $F|K$ (actually it is the maximum). The quoted result refers to the *regular extension* for lower previsions [24]. On the other hand, by considering the *natural extension*, a result equivalent to our Theorem 2 follows [24,16].

Let φ be a 2-monotone capacity on $\mathcal{A}'$ together with its dual $\overline{\varphi}$ and consider

$$\mathcal{P}_\varphi = \{\tilde{P} : \text{probability on } \mathcal{A}' \text{ s.t. } \varphi \leq \tilde{P} \leq \overline{\varphi}\}. \tag{5}$$

If φ is a belief function on a finite Boolean algebra $\mathcal{A}'$, Corollary 3.6 in [14] assures the existence of a finite algebra $\mathcal{A}$ and a probability measure P on $\mathcal{A}$, such that φ is obtained as the lower envelope on $\mathcal{A}'$ of the set of coherent extensions of P on $\mathcal{A} \cup \mathcal{A}'$. In this case, Theorem 2 characterizes the lower envelope of the set of full conditional probabilities obtained extending each $\tilde{P} \in \mathcal{P}_\varphi$ on $\mathcal{A}' \times \mathcal{A}'^0$. Hence, the same theorem characterizes also the lower envelope of the set of coherent extensions on $\mathcal{A}' \times \mathcal{A}'^0$ of a belief function (viewed as a lower probability on $\mathcal{A}'$).

Let $\mathcal{L} = \{H_1, \ldots, H_n\}$ be a finite partition, a *partial prior probability distribution* is a coherent probability P on a set of incompatible events $\{K_1, \ldots, K_m\} \subseteq \langle \mathcal{L} \rangle^0$. In [6] it has been shown that the lower envelope of the set of coherent extensions of P on $\langle \mathcal{L} \rangle$ is a belief function, thus also in this case Theorem 2 characterizes the lower envelope of the coherent extensions on $\langle \mathcal{L} \rangle \times \langle \mathcal{L} \rangle^0$.

5 2-monotone Prior Capacity and Likelihood Function

Given $\mathcal{L}$ and E as in Section 3, here we assume that our knowledge *a priori* is expressed by a 2-monotone capacity φ on $\langle \mathcal{L} \rangle$ while the statistical model is still represented by a likelihood function $f(E|\cdot)$ on $\mathcal{L}$. By Proposition 1 in [18] the assessment $\{\tilde{P}, f\}$ is a coherent conditional probability for every $\tilde{P} \in \mathcal{P}_\varphi$, thus the assessment $\{\varphi, f\}$ is a coherent lower conditional probability. Our aim is to provide a closed form expression for the lower envelope $\underline{P}$ of the set of coherent extensions of $\{\varphi, f\}$ on $\mathcal{A} \times \mathcal{A}^0$, with $\mathcal{A} = \langle \{E\} \cup \mathcal{L} \rangle$.

Next theorem characterizes the lower envelope $\underline{P}(\cdot) = \underline{P}(\cdot|\Omega)$ on $\mathcal{A} \times \{\Omega\}$ as a Choquet integral with respect to φ and it generalizes a result given in [3]. For this aim, for every $F \in \mathcal{A}$ define the $\langle \mathcal{L} \rangle$-measurable function $G_F : \mathcal{L} \to [0,1]$

$$G_F(H_i) = \begin{cases} 0 & \text{if } F \wedge H_i = \emptyset, \\ 1 & \text{if } F \wedge H_i = H_i, \\ f(E|H_i) & \text{if } F \wedge E \wedge H_i \neq \emptyset = F \wedge E^c \wedge H_i, \\ 1 - f(E|H_i) & \text{if } F \wedge E^c \wedge H_i \neq \emptyset = F \wedge E \wedge H_i. \end{cases} \tag{6}$$

Theorem 3. *Given a likelihood function $f(E|\cdot)$ on $\mathcal{L}$ and a 2-monotone capacity $\varphi(\cdot)$ on $\langle \mathcal{L} \rangle$, for every $F \in \mathcal{A}$ it holds*

$$\underline{P}(F) = \oint G_F \mathrm{d}\varphi = \int_0^{+\infty} \varphi\left(\bigvee \{H_i \in \mathcal{L} \,:\, G_F(H_i) \geq x\}\right) \mathrm{d}x.$$

Proof. For every $F \in \mathcal{A}$ and $\tilde{P} \in \mathcal{P}_\varphi$, the probability of F is the expectation of G_F with respect to $\tilde{P}$, so $\underline{P}(F)$ coincides with the minimum of the expectations varying $\tilde{P} \in \mathcal{P}_\varphi$. The proof follows by Proposition 3 in [19] which implies that the lower expectation of G_F with respect to the class of probabilities $\mathcal{P}_\varphi$ coincides with the Choquet integral of G_F with respect to φ. □

Theorem 3 characterizes also the dual upper envelope $\overline{P}(\cdot) = \overline{P}(\cdot|\Omega)$ on $\mathcal{A} \times \{\Omega\}$ as a Choquet integral with respect to $\overline{\varphi}$. Given $\underline{P}(\cdot), \overline{P}(\cdot)$ on $\mathcal{A}$, for every $F|K \in \mathcal{A} \times \mathcal{A}^0$ define

$$L(F \wedge K) = \min\left\{\int G_{F\wedge K} \mathrm{d}\tilde{P} \,:\, \tilde{P} \in \mathcal{P}_\varphi, \int G_{F^c \wedge K} \mathrm{d}\tilde{P} = \overline{P}(F^c \wedge K)\right\}, \tag{7}$$

$$U(F^c \wedge K) = \max\left\{\int G_{F^c\wedge K} \mathrm{d}\tilde{P} \,:\, \tilde{P} \in \mathcal{P}_\varphi, \int G_{F \wedge K} \mathrm{d}\tilde{P} = \underline{P}(F \wedge K)\right\}. \tag{8}$$

Note that it holds in general $\underline{P}(F \wedge K) \leq L(F \wedge K)$ and $U(F^c \wedge K) \leq \overline{P}(F^c \wedge K)$.

The min and max in equations (7) and (8) are attained in correspondence of the extreme points of the set $\mathcal{P}_\varphi$, characterized in [2], whose number is at most $n!$ (i.e., the permutations of $\mathcal{L}$).

Next theorem provides a complete characterization of $\underline{P}(\cdot|\cdot)$ on $\mathcal{A} \times \mathcal{A}^0$ in terms of $\underline{P}(\cdot)$, $\overline{P}(\cdot)$, $L(\cdot)$ and $U(\cdot)$.

Theorem 4. *Given a likelihood function $f(E|\cdot)$ on $\mathcal{L}$ and a 2-monotone capacity φ on $\langle \mathcal{L} \rangle$, for every $F|K \in \mathcal{A} \times \mathcal{A}^0$, $\underline{P}(F|K) = 1$ when $F \wedge K = K$, and if $F \wedge K \neq K$, then:*

(i) if $\underline{P}(K) > 0$ then

$$\underline{P}(F|K) = \min\left\{\frac{\underline{P}(F \wedge K)}{\underline{P}(F \wedge K) + U(F^c \wedge K)}, \frac{L(F \wedge K)}{L(F \wedge K) + \overline{P}(F^c \wedge K)}\right\}; \tag{9}$$

(ii) if $\underline{P}(K) = 0$*, then if* $I \neq \emptyset$ *and* $H_j \wedge F \wedge K \neq \emptyset$ *for all* $j \in J$ *and* $F^c \wedge K \wedge \left(\bigvee_{i\in I} H_i\right)^c = \emptyset$*, where* $I, J \subseteq \{1, \ldots, n\}$ *are, respectively, the maximum and minimum index set such that* $\bigvee_{i\in I} H_i \subseteq K \subseteq \bigvee_{j\in J} H_j$*, then*

$$\underline{P}(F|K) = \min\left\{ \min_{\substack{E\wedge H_i \subseteq F\wedge K \\ E^c\wedge H_i \subseteq F^c \wedge K}} f(E|H_i), \min_{\substack{E^c\wedge H_i \subseteq F\wedge K \\ E\wedge H_i \subseteq F^c \wedge K}} (1 - f(E|H_i)) \right\}; \quad (10)$$

otherwise $\underline{P}(F|K) = 0$*.*

Proof. Let $\mathcal{P} = \{\tilde{P}(\cdot|\cdot)\}$ be the set of full conditional probabilities on $\mathcal{A} \times \mathcal{A}^0$ such that $\tilde{P}_{|\{E\}\times\mathcal{L}} = f$ and $\varphi(\cdot) \leq \tilde{P}(\cdot|\Omega) \leq \overline{\varphi}(\cdot)$, with $\overline{\varphi}$ the dual capacity of φ. If $F \wedge K = K$, then, for every $\tilde{P} \in \mathcal{P}$, it follows $\tilde{P}(F|K) = 1$, which implies $\underline{P}(F|K) = 1$. Hence assume $F \wedge K \neq K$.

To prove condition *(i)*, suppose $\underline{P}(K) > 0$, which implies $\tilde{P}(K) = \tilde{P}(K|\Omega) > 0$ for every $\tilde{P} \in \mathcal{P}$, and so $\underline{P}(F|K) = \min\left\{\frac{\tilde{P}(F\wedge K)}{\tilde{P}(F\wedge K)+\tilde{P}(F^c\wedge K)} : \tilde{P} \in \mathcal{P}\right\}$. The conclusion follows since the real function $\frac{x}{x+y}$ is increasing in x and decreasing in y, so the minimum is attained in correspondence of $\frac{\underline{P}(F\wedge K)}{\underline{P}(F\wedge K)+U(F^c\wedge K)}$ or $\frac{L(F\wedge K)}{L(F\wedge K)+\overline{P}(F^c\wedge K)}$. Finally, condition *(ii)* is implied by the extension procedure described in [8] and Theorem 1. □

In particular, if $\underline{P}(E) > 0$, then for every $F \in \mathcal{A}$ we have $\underline{P}(F\wedge E) = L(F\wedge E)$ and $\overline{P}(F^c\wedge E) = U(F^c\wedge E)$, thus Theorem 4 implies $\underline{P}(F|E) = \frac{\underline{P}(F\wedge E)}{\underline{P}(F\wedge E)+\overline{P}(F^c\wedge E)}$, which coincides with the *lower posterior probability* defined in [25,26].

Note that for all $F|K \in \langle\mathcal{L}\rangle \times \langle\mathcal{L}\rangle^0$, if φ is a belief function and $\underline{P}(K) = \varphi(K) > 0$, then $\underline{P}(\cdot|\cdot)$ on $\langle\mathcal{L}\rangle \times \langle\mathcal{L}\rangle^0$ has the same characterization given in Theorem 2. As a further consequence, for all $F|K \in \mathcal{A} \times \langle\mathcal{L}\rangle^0$, $\underline{P}(F|K)$ can be expressed as the Choquet integral of G_F with respect to the restriction of $\underline{P}(\cdot|K)$ on $\langle\mathcal{L}\rangle$, that is $\underline{P}(F|K) = \oint G_F(\cdot)\mathrm{d}\underline{P}(\cdot|K)$.

Notice that also for the function $\underline{P}(\cdot|K)$ studied in this section (in particular for the lower posterior probability) 2-monotonicity may fail when $\underline{P}(K) = 0$ (see, again, Example 1). In the case the lower posterior probability is 2-monotone, previous results can be used in order to iterate the updating procedure by taking as new prior a lower posterior probability and considering a likelihood function related to another evidence.

References

1. Capotorti, A., Vantaggi, B.: Locally strong coherence in inference processes. Ann. of Math. and Art. Int. 35(1-4), 125–149 (2002)
2. Chateauneuf, A., Jaffray, J.-Y.: Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. Math. Soc. Sci. 17(3), 263–283 (1989)
3. Coletti, G., Gervasi, O., Tasso, S., Vantaggi, B.: Generalized Bayesian inference in a fuzzy context: From theory to a virtual reality application. Comp. Stat. & Data Anal. 56(4), 967–980 (2012)

4. Coletti, G., Petturiti, D., Vantaggi, B.: Possibilistic and probabilistic likelihood functions and their extensions: Common features and specific characteristics. Fuzzy Sets and Sys. (in press), doi: 10.1016/j.fss.2013.09.010
5. Coletti, G., Petturiti, D., Vantaggi, B.: Bayesian inference: the role of coherence to deal with a prior belief function Stat. Meth. and App. (under review)
6. Coletti, G., Scozzafava, R.: Toward a General Theory of Conditional Beliefs. Int. J. of Int. Sys. 21, 229–259 (2006)
7. Coletti, G., Scozzafava, R.: Probabilistic Logic in a Coherent Setting. Trends in Logic, vol. 15. Kluwer Academic Publisher, Dordrecht (2002)
8. Coletti, G., Vantaggi, B.: Probabilistic reasoning with vague information. In: Proc. of the 2nd World Conf. on Soft. Comp., Baku, Azerbaijan, December 3-5 (2012)
9. Császár, Á.: Sur la structure des espaces de probabilitè conditionelle. Acta Math. Ac. Sci. Hung. 6, 337–361 (1955)
10. de Finetti, B.: Sull'impostazione assiomatica del calcolo delle probabilità. Annali Triestini 19(2), 29–81 (1949)
11. de Finetti, B.: Theory of Probability 1-2. John Wiley & Sons, London (1975)
12. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. of Math. Stat. 38(2), 325–339 (1967)
13. Dubins, L.E.: Finitely additive conditional probabilities, conglomerability and disintegrations. Ann. of Prob. 3(1), 89–99 (1975)
14. Fagin, R., Halpern, J.Y.: Uncertainty, belief, and probability. IBM Res. Rep., RJ 6191 (1988)
15. Jaffray, J.-Y.: Bayesian updating and belief functions. IEEE Trans. on Sys., Man and Cyb. 22(5), 1144–1152 (1992)
16. Miranda, E., Montes, I.: Coherent updating of 2-monotone previsions. In: 8th Int. Symp. on Imp. Prob.: Th. and App., Compiègne, France (2013)
17. Regazzini, E.: Finitely additive conditional probabilities. Rend. del Sem. Mat. e Fis. di Milano 55(1), 69–89 (1985)
18. Regazzini, E.: De Finetti's coherence and statistical inference. Ann. of Stat. 15(2), 845–864 (1987)
19. Schmeidler, D.: Integral representation without additivity. Proc. of the Am. Math. Soc. 97(2), 255–261 (1986)
20. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
21. Sundberg, C., Wagner, C.: Generalized Finite Differences and Bayesian Conditioning of Choquet Capacities. Adv. in App. Math. 13(3), 262–272 (1992)
22. Vantaggi, B.: Statistical matching of multiple sources: A look through coherence. Int. J. of Approx. Reas. 49(3), 701–711 (2008)
23. Walley, P.: Coherent lower (and upper) probabilities. Tech. Rep. n. 22, Department of Statistics, University of Warwick (1981)
24. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
25. Wasserman, L.A.: Prior envelopes based on belief functions. Ann. of Stat. 18(1), 454–464 (1990)
26. Wasserman, L.A., Kadane, J.B.: Bayes' theorem for Choquet capacities. Ann. of Stat. 18(3), 1328–1339 (1990)
27. Williams, P.M.: Note on conditional previsions. Int. J. of Approx. Reas. 44, 366–383 (2007); (Unpublished manuscript 1975)

Forecasting Short Time Series with the Bayesian Autoregression and the Soft Computing Prior Information

Olgierd Hryniewicz and Katarzyna Kaczmarek

Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw, Poland
{Olgierd.Hryniewicz,K.Kaczmarek}@ibspan.waw.pl

Abstract. As observed in a real-life production company, there is often a need to forecast demand for new products, despite the shortness of the available time series data. We introduce an innovative approach to discover prior information from experts in their fields and incorporate this into the bayesian autoregressive forecasting. It is observed that for the short time series, the bayesian method combined with the soft computing techniques, especially the linguistic summarization and the supervised learning, outperforms the traditional, statistical methods, and that prior assumptions play a key role. The details of the proposed approach are illustrated by the simulation study.

Keywords: Time Series Forecasting, Autoregression, Bayesian Methods, Data Analysis, Summarization, Decision Support, Classification, Support Vector Machine.

1 Introduction

When forecasting demand for a new product or a new customer, there are usually very few time series observations available. For such short time series the traditional methods for the time series forecasting may be inaccurate. Hopefully, there are analysts and experts that conduct the forecasting process based on their expertise and intuitions. However, the research on the psychology of decision-making proves that people may take irrational decisions. Therefore, our main objective is to create a human-consistent tool to support the forecasting of the short autoregressive time series.

The focus of this paper is to present the *Bayesian Forecasting with Soft Computing Prior Information* (BFSC) approach that supports the decision-making about the prior assumptions for the short autoregressive time series. We process the linguistic summaries, which are a result of data mining and are easily interpretable for experts.

Within the experiments the significance of the prior assumptions is verified and the comparative analysis of the forecasting accuracy with the well-known forecasting methods is performed.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_10

The structure of this paper is as follows. The next section explains the motivation and background for the problem. In Section 3 we present the details of the proposed approach for forecasting. The performance of the approach is illustrated with the simulation studies in Section 4. This paper concludes in Section 5 with the summary and the potential further research opportunities.

2 Background and Motivation

The Box-Jenkins autoregressive and moving average (ARMA) processes are one of the most popular probabilistic models for forecasting, and although simple, very successful in applications. However, the traditional parameter estimation methods, like the Yule-Walker or the Burg algorithm, require the availability of the time series history of at least 50 observations (see Box [1]). If a time series is short, then the estimation methods may not be accurate. For the review of methods to estimate the autoregressive parameters, refer to [2].

Hopefully, the bayesian methods enable the inclusion of the prior information with promising results for the short time series. However, the proper selection of prior probability distributions for the unknown variables is essential for the satisfactory forecasting performance in terms of accuracy and time. Following [3], the definitions for the prior probability distributions are usually assumed subjectively based on expert's experience. In [4], the authors show the critical importance of the prior assumptions for the bayesian model averaging. Unfortunately, the research on the behavioral economics and the psychology of decision-making proves that people may be misled by emotions, lack of skills, extensive self-confidence or risk avoidance. Therefore, there is a need to support the decision-making about the prior assumptions.

The soft computing research for time series has gained a lot of attention in literature over the last decade. For the recent review on the time series data mining see [5]. One of the goals of the data mining is to provide the human-consistent description of the raw data. Linguistic summaries in the sense of [6] are an example of such descriptions. The linguistic summaries are in line with the visual inspection capabilities and describe general facts about the evolution of a time series in a (quasi) natural language e.g., *'Among all increasing trends, majority is long.'*, and therefore, are easily interpretable for experts.

Within this paper, the linguistic summarization and the supervised learning are applied to discover the information about the expected trends in time series. Then, the data mining results are included as the prior information in the process of the bayesian autoregressive forecasting.

3 Proposed Algorithm

We introduce the *Bayesian Forecasting with Soft Computing Prior Information* (BFSC) method. Its main objective is to construct the prior model probabilities combining the data mining results and the experts' advice.

The method consists of the following three steps: the supervised learning of the probabilistic models, the mining for the human-consistent prior information, and finally, the bayesian posterior simulation and forecasting.

The input data for the algorithm are as follows:

- Time series for prediction $y = \{y_t\}_{t=1}^{n}$, where $n \in \{n_{min}, ..., n_{max}\} \subseteq \mathbb{N}$, $y \in \mathbb{Y}$, where $\mathbb{Y}$ is a space of discrete time series
- The selection of the template probabilistic models $M = \{M_1, M_2, ..., M_J\} \subseteq \mathbb{M}$ where $\mathbb{M}$ is a set of stationary autoregressive processes. For $i \in \{1, ..., J\}$: $\theta_{k_i} \in \Theta_M \subseteq \mathbb{R}^{k_i}$ is the vector of unobservables (parameters) for the autoregressive process M_i.

The proposed algorithm consists of the following steps:

1. **The supervised learning of the probabilistic models**
   The goal is to build the training database and to discover rules enabling the classification of the probabilistic models based on the sets of linguistic summaries describing the evolution of time series.
   1.1. **Generate time series realizations of the template models**
   Generate the training database from M: $T_m^s = \{\{y_t^1\}_{t=1}^{m}, ..., \{y_t^s\}_{t=1}^{m}\} \subseteq \mathbb{Y}^s$, where $m \in \mathbb{N}$ and $s = J \times k$ with $(k \geq 10) \wedge (k \in \mathbb{N})$. Create $C^s = \{c(y^1), ..., c(y^s)\}$, where $c : \mathbb{Y} \to M$ assigns to the time series its model.
   1.2. **Perform segmentation**
   Transform the time series T_m^s into the series of meaningful labeled intervals (trends) $Tr^s = \{\{Tr_t^1\}_{t=1}^{m_1}, ..., \{Tr_t^s\}_{t=1}^{m_s}\}$. We adapt a broken-line segmentation algorithm based on the idea of a sliding window.
   1.3. **Summarize series of trends**
   Discover linguistic summaries $LI_{Tr}^s = \{\{Li_h^1\}_{h=1}^{z}, ..., \{Li_h^s\}_{h=1}^{z}\} \subseteq \mathbb{LI}^s$ where linguistic summary LI is defined according to the classic calculus of linguistically quantified proposition based on the concept of the Yager's extended protoform [6]. The linguistic summary LI : Q R *trends are* P consists of quantifier Q (e.g., *most, among all*), qualifier R (attribute together with an imprecise label, e.g., *increasing trend*), summarizer P (attribute together with an imprecise label). For details see also [7].
   1.4. **Evaluate the quality of the linguistic summaries**
   Measure the quality of linguistic summaries $LI \in LI_{Tr}^s$ and save as $V^s = \{\{V_h^1\}_{h=1}^{z}, ..., \{V_h^s\}_{h=1}^{z}\}$ where $V : LI \to [0, 1]$. We adapt the degree of truth (validity) V introduced by Zadeh and defined as follows:

$$V(Q\ R\ trends\ are\ P) = \mu_Q(\frac{\sum_{i=1}^{n} (\mu_R(y_n) \wedge \mu_P(y_n))}{\sum_{i=1}^{n} \mu_R(y_n)}) \tag{1}$$

   where $\mu_R(y_n), \mu_P(y_n)$ are the membership functions $\mu_R, \mu_P : \mathbb{R} \to [0, 1]$ determining the degree to which R, P respectively, are satisfied for the time series y at the given moment n. For basic definitions related to the fuzzy sets theory see e.g., [8].

1.5. **Apply Support Vector Machines (SVM)**
Learn the classification rules on vectors V^s and C^s from the training database. The elements of V^s are attributes for the classification task. The elements of C^s, that correspond to the template probabilistic models, are the classes (categories). For details on the supervised learning methods and the SVM classification see e.g., [9,10].

2. **The mining for human-consistent prior information**
The goal is to discover the linguistic summaries about the expected evolution of the predicted time series from the experts of the field.
2.1. **Generate provisional linguistic summaries**
Create the set of the expected linguistic summaries LI^E related to the short time series y for prediction and estimate the vector $V^E \in [0,1]^E$ with the respective degrees of validity.
2.2. **Validate linguistic summaries with experts**
Present V^E to an expert for the evaluation through the human-computer interaction.
2.3. **Calculate the prior information**
Calculate the classification scores $Sc(V^E) = \{Sc^{M_1}, Sc^{M_1}, ..., Sc^{M_J}\}$ where $Sc : [0,1]^E \to [0,1]^J$ using rules learned through the SVM in step (1.5). The classification assumes assigning score to each of J possible categories (models).

3. **The posterior simulation and forecasting**
The goal is to find forecast $\omega = \{y_{n+l}\}$ *being the vector of interest for the bayesian inference.*
3.1. **Construct the prior probability distributions**
For all $k \in \{1, ..., s\}$ establish $p(M_k|M)$ based on the classification scores $Sc(V^E)$. Let r denote the acceptance rate, if $Sc^{M_k} > r$, then include M_k in the posterior simulation and add k to J_r. The priors for the unobservables $p(\theta_k|M_k)$ are adapted from the definitions related to the template models $p(M_k|M)$.
3.2. **Apply the bayesian averaging**
Include the multiple models in the process of the inference

$$p(\omega|y, M) = \sum_{j=1}^{J_r} p(\omega|y, M_j)p(M_j|M) \tag{2}$$

where $p(\omega|y, M_j)$ is the posterior density of the vector of interest.
3.3. **Run the Markov Chain Monte Carlo posterior simulation**
Generate forecasts for the time series y. MCMC yields a pseudo-random sequence of the vector of interest to estimate its posterior moments concluding from the Bayes Theorem according to the following formula:

$$p(\theta_M|y, M) = \frac{p(\theta_M|M)p(y|\theta_M, M)}{p(y|M)} \tag{3}$$

instead of describing the probability density function analytically. For details see e.g., [3].

3.4. **Evaluate the quality of the forecast**
For example, for the normalized time series use the value of sum of squared errors (SSE) to measure the forecasting accuracy. The measures for the forecasting accuracy are studied by [11].

4 Experimental Results

This section illustrates the performance of the proposed *Bayesian Forecasting with Soft Computing Prior Information* (BFSC) approach and its forecasting accuracy. The time series datasets for the simulation study are created from the stationary autoregressive processes in line with the following formula:

$$\tilde{y_t} = \phi_1 \tilde{y_{t-1}} + a_t,\ a_t\ \sim N(0, \sigma^2),\ \tilde{y_t} = y_t - \mu \tag{4}$$

where $\theta = \{\phi_1, \sigma^2\}$ is the vector of the unknown variables, $\phi_1 \in [-0.9, 0.9], \sigma^2 \in [0, 1]$ and μ is the mean of y_t.

The program with the proposed approach is created in Python with the support of NumPy, SciPy, PyMc extension modules. Linguistic summaries are generated with the Trend Analysis System [12].

The performance of the proposed approach is compared to the alternative bayesian least squares method with the uninformative prior information and to the two well-known, popular in applications, non-bayesian estimation methods: the Yule-Walker and the Burg method.

4.1 Supervised Learning of Probabilistic Models

First, we analyze the classification performance for the database with the template time series and the linguistic summaries used to learn the system. It is observed that the degree of truth for almost half of all the linguistic summaries that were generated automatically for the template time series is close to 0.

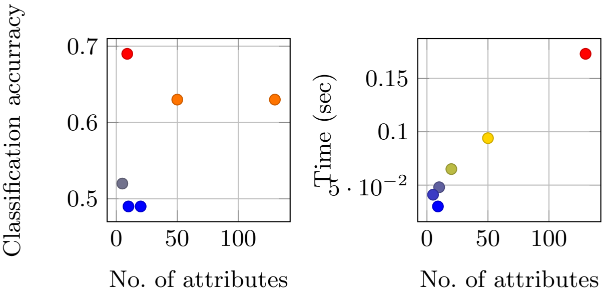


Fig. 1. Classification accuracy and time for different number of attributes (linguistic summaries) for the 3-class problem, SVM with RBF kernel

We examine the classification accuracy and time depending on the considered number of attributes (features) and classes (models) based on the mean for 5-fold cross validation. As demonstrated by results for the 3-class problem in Figure 1, the classification accuracy is highest and amounts to 0.69, 0.63 and 0.63 for 9, 50 and 130 attributes, respectively. The scenario of 9 attributes refers to the set of simple linguistic summaries based on the short protoforms. Classification time increases proportionally to the size of the considered attributes. We conclude that the subspace of attributes for classification (linguistic summaries) may be limited leading to the increased efficiency and without the loss on the accuracy. Furthermore, the limited subspace is easier for the experts' interpretation.

4.2 Mining for the Human-consistent Prior Information

The attributes (and labels) of trends for the time series segmentation and summarization considered in the experiment are as follows: duration (*short, medium, long*), dynamics (*increasing, constant, decreasing*), variability (*low, moderate, high*). The linguistic summaries are naturally linked to the expressions in natural language, and therefore, are easily interpreted and validated by humans.

Table 1. Avg degree of truth V for the selected linguistic summaries in the groups of time series generated from 3 models: M_1, M_2, M_3 with autoregressive coefficient $\phi_1 = 0.0, \phi_1 = 0.5, \phi_1 = 0.9$, respectively

Description of the linguistic summary	V		
	M_1	M_2	M_3
Among all trends, most are short	1.00	0.71	0.49
Among all trends, most are moderate	0.47	0.81	0.86
Among all trends, most are medium	0.51	0.88	0.98
Among all trends, most are low	0.47	0.22	0.14
Among all decr trends, most are medium	0.17	0.46	0.69
Among all decr trends, most are moderate	0.34	0.44	0.71

Table 1 presents average values of the degree of validity V for groups of the sample time series generated from the selected 3 template models. The degree of validity V is a satisfactory discrimination between the considered probabilistic models.

4.3 Posterior Simulation and Forecasting

As presented in Table 2, the forecasting accuracy measured by the sum of squared errors for the 1-step-ahead predictions of the proposed bayesian approach with the soft computing prior outperforms significantly the alternative methods for the short time series (10-20 observations).

Table 2. The forecasting accuracy measured by the sum of squared errors for 1-step-ahead prediction on the samples of 100 time series for: a) **BFSC**: the proposed approach b) **B-un**: the bayesian least squares methods with uninformative prior c) **Burg**: the Burg method d) **Y-W**: the Yule Walker method

	No of obs in TS												
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**50**	**100**
BFSC	107.3	107.2	107.4	107.3	106.9	108.5	106.3	105.8	106.8	106.6	106.8	103.2	103.0
B-un	123.1	119.0	118.6	121.2	120.6	117.8	115.5	114.7	114.8	112.1	113.2	106.1	104.0
Burg	118.2	116.7	117.4	116.2	116.0	114.7	113.5	111.4	111.2	111.5	111.8	105.2	104.3
Y-W	119.2	117.6	117.5	116.2	116.6	114.9	113.9	112.9	112.6	113.0	112.8	105.0	104.0

At the same time, it is observed that for longer time series (100 observations), there is no significance difference between the SSE obtained by the considered methods in terms of the forecast accuracy.

Table 3 shows the efficiency of the estimation measured by the sum of squared error (SSE) of posterior mean of the autoregressive coefficient compared to the true coefficient which was assumed to generate the time series. As presented in Table 3, the accuracy of parameter estimation depends on the length of the available time series data and is 2 to 4 times more accurate for the proposed approach than for the alternative methods. As to consider an example, for the time series of length 10, the SSE of the posterior mean amounts to 1.95, 8.42, 8.54 and 9.85 respectively, for the proposed BFSC, the Burg algorithm, the Yule-Walker method and the bayesian least squares with uninformative prior. For time series of length 20, within this dataset, the SSE of the posterior mean amounts to 1.63, 4.12, 4.32 and 4.92, respectively. Similarly to the forecast accuracy, the SSE of the posterior mean is comparable for all methods and ranges between 0.58 to 0.81.

Table 3. The efficiency of estimation measured by the sum of squared error of the posterior mean for the autoregressive coefficient compared to the coefficient which was assumed to generate the time series on the samples of 100 time series for: a) **BFSC**: the proposed approach b) **B-un**: the bayesian least squares methods with uninformative prior c) **Burg**: the Burg method d) **Y-W**: the Yule Walker method

	No of obs in TS													
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**50**	**100**	**Avg**
BFSC	1.95	1.84	1.92	1.78	1.81	1.74	1.65	1.69	1.76	1.61	1.63	0.96	0.58	1.61
B-un	9.85	9.63	7.89	7.82	7.51	7.12	5.69	5.93	5.95	5.31	4.92	1.58	0.81	6.15
Burg	8.42	8.41	7.10	6.60	5.96	5.55	4.95	5.03	4.68	4.25	4.12	1.41	0.74	5.17
Y-W	8.54	8.37	7.31	6.69	6.21	5.66	5.10	5.23	4.84	4.46	4.32	1.44	0.75	5.30

The performed experiment confirmed that the proposed approach helps to increase the accuracy of forecasting time series with 10-20 observations.

5 Conclusion

In this paper we have introduced the *Bayesian Forecasting with Soft Computing Prior Information* approach to support the decision-making about the prior assumptions for forecasting. The simulation of the forecast accuracy showed that for the short autoregressive time series (10-20 observations) the proposed method may lead to the increase of the forecasts accuracy compared to the traditional estimation methods. Because of the purely simulational character of this study, the obtained results are of preliminary character, and in future should be confirmed by the experiments performed on real-life data sets.

The approach may be easily extended with other template probabilistic models. However, further experiments are needed to prove its overall effectiveness. For future research we plan to include multivariate linguistic summaries, multiple interpretations of the processed imprecise labels and other types of the linguistic information like frequent patterns and association rules.

Acknowledgments. The authors thank Janusz Kacprzyk and Anna Wilbik for their support and advice concerning the Trend Analysis System.

Katarzyna Kaczmarek is supported by the Foundation for Polish Science under International PhD Projects in Intelligent Computing. Project financed from the European Union within the Innovative Economy Operational Programme 2007-2013 and European Regional Development Fund.

References

1. Box, G., Jenkins, G.: Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1970)
2. Monson, H.: Statistical Digital Signal Processing and Modeling. John Wiley & Sons (1996)
3. Geweke, J.: Contemporary Bayesian econometrics and statistics. In: Wiley Series in Probability and Statistics. John Wiley (2005)
4. Ley, E., Steel, M.: On the effect of prior assumptions in Bayesian Model Averaging with applications to growth regression. Journal of Applied Econometrics 24, 651–674 (2009)
5. Fu, T.: A review on time series data mining. Engineering Applications of Artificial Intelligence 24, 164–181 (2011)
6. Yager, R.: A new approach to the summarization of data. Inf. Sci. 28, 69–86 (1982)
7. Kacprzyk, J., Zadrożny, S.: Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. Information Sciences 173, 281–304 (2005)
8. Gil, M., Hryniewicz, O.: Statistics with imprecise data. In: Encyclopedia of Complexity and Systems Science, pp. 8679–8690. Springer, Heidelberg (2009)
9. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
10. Berthold, M., Hand, D.: Intelligent data analysis. An Introduction. Springer (2007)
11. Hyndman, R., Koehler, A.: Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688 (2006)
12. Kacprzyk, J., Wilbik, A., Partyka, A., Ziólkowski, A.: Trend Analysis System, Systems Research Institute of Polish Academy of Sciences, Warsaw (2011)

Learning Structure of Bayesian Networks by Using Possibilistic Upper Entropy

Mathieu Serrurier[1] and Henri Prade[1,2]

[1] IRIT - University Paul Sabatier, Toulouse, France
serrurie@irit.fr
[2] QCIS, University of Technology, Sydney, Australia
prade@irit.fr

Abstract. The most common way to learn the structure of Bayesian networks is to use a score function together with an optimization process. When no prior knowledge is available over the structure, score functions based on information theory are used to balance the entropy of the conditional probability tables with network complexity. Clearly, this complexity has a high impact on the uncertainty about the estimation of the conditional distributions. However, this complexity is estimated independently of the computation of the entropy and thus does not faithfully handle the uncertainty about the estimation. In this paper we propose a new entropy function based on a "possibilistic upper entropy" which relies on the entropy of a possibility distribution that encodes an upper bound of the estimation of the frequencies. Since the network structure has a direct effect on the number of pieces of data available for probability estimation, the possibilistic upper entropy is of an effective interest for learning the structure of the network. We also show that possibilistic upper entropy can be used for obtaining an incremental algorithm for the online learning of Bayesian network.

1 Introduction

Bayesian networks [8] are compact representations of probabilistic dependencies over a set of variables. A Bayesian networks (BN) is composed of a directed acyclic graph (DAG) which encodes the dependency relations, and of tables which describe the conditional probability distributions. Given a DAG and a set of complete vectors over variables, the tables can be easily obtained by computing conditional frequencies (which can be refined with a smoothing process). Thus, given a set of complete vectors over variables, a challenge is to identify the best structure for the BN. The best structure is theoretically the one in which the entropy of the conditional probabilities is the lowest. However, adding an edge (and then a dependency) in the graph always decreases the entropy, but it also decreases the amount of data used for estimating the conditional probability distributions. Learning the structure of a BN thus consists in finding the best trade-off between the global entropy of the BN and the uncertainty around the estimation of the conditional probabilities. Since the uncertainty is related to the complexity of the DAG (i.e. the size of the tables), score functions based on

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_11

information theory, such as Akaike information criterion (AIC) or minimum description length (MDL), have been currently used (other measures based on prior knowledge over structure have also be proposed in [7], but we do not consider them in this paper since we assume that we have no prior knowledge). These score functions balance the entropy values with the complexity of the graph. Their major limitation is that they consider the computations of entropy and of structure complexity in an independent way. Thus, it does not reflect the manner how the information is dispatched in the table. In this paper we propose to use the upper bound of the frequency estimates for defining a so-called possibilistic upper entropy (π-*up* entropy). The approach relies on the building of a possibility distribution. Quantitative possibility measures can be viewed as upper bounds of probabilities. Then, a possibility distribution represents a family of probability distributions [5]. This view was first implicitly suggested in [10] when emphasizing the idea that what is probable must be possible. Following this intuition, a probability-possibility transformation has been proposed [6]. This transformation associates a probability distribution with the maximally specific (restrictive) possibility distribution which is such that the possibility of any event is an upper bound of the corresponding probability. Possibility distributions are then able to describe epistemic uncertainty and to represent knowledge states such as total ignorance, partial ignorance, or complete knowledge. In the spirit of [9], we propose a log-based loss function for possibility distributions. We derive an entropy function for a possibility distribution associated to a frequency distribution. In order to obtain the π-*up* entropy for a frequency distribution, we build a possibility distribution that upper bounds the confidence intervals of the frequency values (according to the amount of data available and a confidence degree) and we compute its relative possibilistic entropy. This π-*up* entropy has a nice behavior. For instance, it respects the entropy order for a fixed level of information and it increases the entropy value for a fixed frequency distribution when the amount of data decreases. Our π-*up* entropy shares similar ideas (handling the uncertainty around the estimation of the probability values) with a proposal by Abelan *et al.* [1] for credal sets. Our approach is simpler and easier to compute. Their entropy function is based on the worst entropy value for the probabilities in the credal set obtained by the computation of confidence intervals. Thus, in order to have discriminant values, they have to use very optimistic confidence intervals (while we compute faithful confidence intervals). Moreover, the computation of entropy based on credal sets requires the solving of a simplex problem and would make this approach time consuming.

In this paper, we show that we can directly use π-*up* entropy as a score function for learning the structure of Bayesian networks. In addition to the classical learning approach based on optimization, we propose a very simple incremental learning method. The paper is organized as follows. First we provide a short background on possibility distributions and possibility measures and their use as upper bound of families of probability distributions. Second, we describe probabilistic entropy and π-*up* entropy and their properties. Section 4 is devoted to the presentation of the algorithms for learning the structure of BN's. In the last section, we compare our score function with state of the art ones on 10 benchmark databases, which shows a clear benefit for the approach.

2 Possibility Theory

Possibility theory, introduced in [10], was initially proposed in order to deal with imprecision and uncertainty due to incomplete information as the one provided by linguistic statements. This kind of epistemic uncertainty cannot be handled by a single probability distribution, especially when a priori knowledge about the nature of the probability distribution is lacking. A possibility distribution π is a mapping from Ω to $[0, 1]$. We only consider the case where $\Omega = \{C_1, \ldots, C_q\}$ is a discrete universe (of classes in this paper). The value $\pi(x)$ denotes the possibility degree of x. For any subset of Ω, the possibility measure is defined as follows :

$$\forall A \in 2^{\Omega}, \Pi(A) = max\{\pi(x), x \in A\}.$$

If it exists at least one singleton $x \in \Omega$ for which we have $\pi(x) = 1$, the distribution is normalized. We can distinguish two extreme cases of knowledge situation: complete knowledge when $\exists x \in \Omega$ such as $\pi(x) = 1$ and $\forall y \in \Omega, y \neq x, \pi(y) = 0$ and total ignorance when $\forall x \in \Omega, \pi(x) = 1$.

The natural pre-order over possibility distributions (named *specificity*) is defined by the classical function pre-order. Namely, a distribution π is more specific than π', denoted $\pi \preceq \pi'$, if and only if $\forall x \in \Omega, \pi(x) \leq \pi'(x) \Leftrightarrow \forall A \in 2^{\Omega}, \Pi(A) \leq \Pi'(A)$.

One view of possibility theory is to consider a possibility distribution as a family of probability distributions (see [3] for an overview). Thus, a possibility distribution π will represent the family of the probability distributions for which the measure of each subset of Ω will be respectively lower and upper bounded by its necessity and its possibility measures. More formally, if $\mathcal{P}$ is the set of all probability distributions defined on Ω, the family of probability distributions $\mathcal{P}(\pi)$ associated with π is defined as $\mathcal{P}(\pi) = \{p \in \mathcal{P}, \forall A \in \Omega, N(A) \leq P(A) \leq \Pi(A)\}$, where P is the probability measure associated with p. In this scope, the situation of total ignorance corresponds to the case where all probability distributions are possible. According to this probabilistic interpretation, Dubois *et al.* [6] propose to transform a probability distribution into a possibility distribution by choosing the most informative possibility measure that upper bounds the considered probability measure. This possibility measure corresponds to the tightest possibility distribution. Let us consider a probability distribution p on $\Omega = \{C_1, \ldots, C_q\}$. We note $\sigma \in S_q$ a permutation of the set $1, \ldots, q$. For each permutation $\sigma \in S_q$ we can build a possibility distribution π_p^{σ} which encodes p as follows:

$$\forall j \in \{1, \ldots, q\}, \pi_p^{\sigma}(C_j) = \sum_{k, \sigma(k) \leq \sigma(j)} p(C_k). \tag{1}$$

Then, each π_p^{σ} corresponds to a cumulative distribution of p according to the order defined by σ. We have $\forall \sigma \in S_q, p \in \mathcal{P}(\pi_p^{\sigma})$. The probability-possibility transformation [4] uses one of these particular possibility distributions. Given a probability distribution p on $\Omega = \{C_1, \dots, C_q\}$ and a permutation $\sigma^* \in S_q$ such as $p(C_{\sigma^*(1)}) \leq \dots \leq p(C_{\sigma^*(q)})$, the probability possibility of p is noted π_p^* and is defined as $\pi_p^* = \pi_p^{\sigma^*}$. π_p^* is the cumulative distribution of p built by considering the increasing order of p. For this order, π_p^* is the most specific possibility distribution that encodes p.

3 Possibilistic Upper Entropy

In section we explain how particular possibility distributions can be used to take into account the amount of data used for estimating the frequencies into the computation of the entropy. Probabilistic loss functions are used for evaluating the adequateness of a probability distribution with respect to data. We consider a set of realizations $X = \{x_1, \dots, x_n\}$ of a random variable over a discrete universe $\Omega = \{C_1, \dots, C_q\}$. Let $\alpha_1, \dots, \alpha_q$ be the frequency of the elements of X that belong respectively to $\{C_1, \dots, C_q\}$. The log-likelihood is a natural loss function for estimating the adequateness between a probability distribution p on the discrete space $\Omega = \{C_1, \dots, C_q\}$ and an event x_i. Formally the likelihood coincides with a probability value. The logarithmic-based likelihood is defined as follows:

$$\mathcal{L}_{log}(p|x_i) = -\sum_{j=1}^{q} \mathbb{1}_j(x_i) log(p(C_j)), \tag{2}$$

where $\mathbb{1}_j(x_i) = 1$ if $x_i = C_j$, and $\mathbb{1}_j(x_i) = 0$ otherwise. When we consider the whole set of data we obtain $\mathcal{L}_{log}(p|X) = -\sum_{j=1}^{q} \alpha_j log(p(C_j))$. When p is estimated with respect to frequencies, we obtain the entropy of the distribution.

$$\mathcal{H}(p) = -\sum_{j=1}^{q} p(C_j) log(p(C_j)). \tag{3}$$

The entropy measures the amount of information of the distribution. The higher the entropy, the lower the amount of information (uniform distribution). We now show how to use $\mathcal{L}_{log}$ in order to define a loss function, and the related entropy, for possibility distributions that agree with the interpretation of a possibility distribution in terms of a family of probability distributions. Proofs and detailed discussion about possibilistic loss function can be found in [9]. We expect three properties:

(a) the loss function is minimal for the possibility distribution that results from the probability-possibility transformation of the frequencies
(b) the possibilistic entropy is the sum of the independent loss functions for each event as for probabilistic entropy
(c) the possibilistic entropy of the results of the probability-possibility transformations agree with the probabilistic entropy order.

Since a possibility distribution π can be viewed as an upper bound of a cumulative function, for all j, the pair $\pi_j = (\pi(C_{\sigma(j)}), 1-\pi(C_{\sigma(j)}))$ (σ is the permutation of S_q such that $\pi(C_{\sigma(1)}) \leq \ldots \leq \pi(C_{\sigma(q)})$) can be seen as a binomial probability distribution for the sets of events $BC_j = \bigcup_{i=1}^{j} C_{\sigma(i)}$ and $\overline{BC_j}$. Then, the logarithmic loss of a possibility distribution for an event will be the average of the log loss of each binomial distribution π_j.

$$\mathcal{L}_{pos}(\pi|x_i) = \frac{\sum_{j=1}^{q} \mathcal{L}_{log}(\pi_j|x_i)}{q} \tag{4}$$

When we consider the whole set of data, we obtain:

$$\mathcal{L}_{pos}(\pi|X) = -\frac{\sum_{j=1}^{q}(cdf_j * log(\pi(C_j) + (1 - cdf_j) * log(1 - \pi(C_j)))}{q} \tag{5}$$

where $cdf_j = \sum_{k,\sigma(k)\leq\sigma(j)} \alpha_k$. The property (a) has been proven in [9]. We remark that cdf_j corresponds to the cumulative probability distribution of the frequencies with respect to σ (Eq. 1). Then, we can derive a definition of the entropy of a possibility distribution π relative to a probability distribution p by considering the cumulative distribution of p according to the order σ (π_p^σ):

$$\mathcal{H}_{pos}(p,\pi) = -\frac{\sum_{j=1}^{q} \pi_p^\sigma(C_j) * log(\pi(C_j))}{q} - \frac{\sum_{j=1}^{q}(1 - \pi_p^\sigma(C_j)) * log(1 - \pi(C_j))}{q} \tag{6}$$

The expected property (b) is obvious if we consider the probability distribution p such as $p(C_i) = \alpha_i$. We can establish some properties of possibilistic entropy which validate the property (c) and show that the possibility entropy is fully compatible with the interpretation of a possibility distribution as a family of probability distributions:

- Given two probability distributions p and p' on $\Omega = \{C_1, \ldots, C_q\}$ we have $\mathcal{H}(p) \leq \mathcal{H}(p') \Rightarrow \mathcal{H}_{pos}(p, \pi_p^*) \leq \mathcal{H}_{pos}(p', \pi_{p'}^*)$,
- Given a probability distribution p and two possibility distributions π and π' on $\Omega = \{C_1, \ldots, C_q\}$ we have $\pi_p^* \preceq \pi \preceq \pi' \Rightarrow \mathcal{H}_{pos}(p, \pi_p^*) \leq \mathcal{H}_{pos}(p, \pi) \leq \mathcal{H}_{pos}(p, \pi')$.

As said previously, the entropy calculus does not take into account the amount of information used for estimating the frequencies. The idea behind π-*up* entropy is to consider the confidence intervals around the estimation of the frequencies to have an entropy measure that increases when the size of the confidence interval increases. Applying directly the entropy to the upper-bounds of the frequency is not satisfactory since entropy only applies to genuine probability distributions. Similarly, using the probability distribution that has values in the confidence interval and that has the maximum value of entropy is too restrictive. Thus we propose to build the most specific possibility distribution that upper bounds the confidence interval and compute its possibilistic entropy relative to the frequency distribution.

We use the Agresti-Coull interval (see [2] for a review of confidence intervals for binomial distributions) for computing the upper bound value of the probability of an event. Given $p(c)$ the probability of the event estimated from n pieces of data, the upper bound $p^*_{\gamma,n}$ of the $(1-\gamma)\%$ confidence interval of the distribution is obtained as follows:

$$p^*_{\gamma,n}(c) = \tilde{p} + z\sqrt{\frac{1}{\tilde{n}}\tilde{p}(1-\tilde{p})} \tag{7}$$

where $\tilde{n} = n + z^2$, $\tilde{p} = \frac{1}{\tilde{n}}(p(c) * n + \frac{1}{2}z^2)$, and z is the $1 - \frac{1}{2}\gamma$ percentile of a standard normal distribution. The most specific $\pi^\gamma_{p,n}$ that upper bounds the $(1-\gamma)\%$ confidence interval of the probability distribution p on $\Omega = \{C_1, \ldots, C_q\}$ estimated from n pieces of data is computed as $\pi^\gamma_{p,n}(C_j) = P^*_{\gamma,n}(\bigcup_{i=1}^{j} C_{\sigma(i)})$ where $\sigma \in S_q$ is the permutation such as $p(C_{\sigma(1)}) \leq \ldots \leq p(C_{\sigma(q)})$. Then $\pi^\gamma_{p,n}$ is built in the same way as π^*_p except that it also takes into account the uncertainty around the estimation of p. Obviously, we have $p \in \mathcal{P}(\pi^\gamma_{p,n})$, $\forall n > 0, \pi^*_p \preceq \pi^\gamma_{p,n}$ and $\lim_{n\to\infty} \pi^\gamma_{p,n} = \pi^*_p$. Having $\pi^\gamma_{p,n}$, we can now define the *π-up* entropy of a probability distribution:

$$\mathcal{H}_{\pi\text{-}up}(p, n, \gamma) = \mathcal{H}_{poss}(p, \pi^\gamma_{p,n}) \tag{8}$$

$\mathcal{H}_{\pi\text{-}up}$ has the following properties:

- Given a probability distribution p on $\Omega = \{C_1, \ldots, C_q\}$ and $n' \leq n$ we have $\forall \gamma \in]0, 1[, \mathcal{H}_{\pi\text{-}up}(p, n, \gamma) \leq \mathcal{H}_{\pi\text{-}up}(p, n', \gamma)$,
- Given two probability distributions p and p' on $\Omega = \{C_1, \ldots, C_q\}$ we have $\forall \gamma \in]0, 1[, \mathcal{H}(p) \leq \mathcal{H}(p') \Rightarrow \mathcal{H}_{\pi\text{-}up}(p, n, \gamma) \leq \mathcal{H}_{\pi\text{-}up}(p', n, \gamma)$.

4 Learning a Bayesian Network Structure

We consider a BN over a set of m random variables $\mathcal{V} = \{V_1, \ldots, V_m\}$ (each random variable V_i can take r_i possible values). $\mathcal{D}$ is a set of n complete valuations of $\mathcal{V}$. Given a Bayesian network $\mathcal{B}$, we note q_i the numbers of lines in the conditional table for the variable V_i. Given $\mathcal{B}$ and $\mathcal{D}$ we define the *AIC* and *MDL* score functions as follows:

$$AIC(\mathcal{B}, \mathcal{D}) = LogP(\mathcal{B}, \mathcal{D}) - Dim(\mathcal{B}), \tag{9}$$

$$MDL(\mathcal{B}, \mathcal{D}) = LogP(\mathcal{B}, \mathcal{D}) - \frac{1}{2}Dim(\mathcal{B}) * log(n), \tag{10}$$

where $Dim(\mathcal{B}) = \sum_{i=1}^{m}(1 - r_i) * q_i$. The terms $LogP(\mathcal{B}, \mathcal{D})$ is closely related to the entropy of the conditional distribution (thanks to the decomposability of the entropy) when they are evaluated by considering the frequencies:

$$LogP(\mathcal{B}, \mathcal{D}) = \sum_{i=1}^{m}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{i,j,k} * log(\frac{N_{i,j,k}}{N_{i,j}}) = -\sum_{i=1}^{m}\sum_{j=1}^{q_i} N_{i,j} * \mathcal{H}(p_{i,j}) \tag{11}$$

where $N_{i,j,k}$ is the number of examples in $\mathcal{D}$ which fall in the jth line of the table of V_i and for which V_i takes the kth possible value, $N_{i,j} = \sum_{k=1}^{r_i} N_{i,j,k}$, and $p_{i,j}$ is the conditional probability distribution in the jth line of the table of V_i. Since $\mathcal{H}_{\pi\text{-}up}$ is also decomposable, we propose the following score function

$$POSS(\mathcal{B}, \mathcal{D}) = -\sum_{i=1}^{m}\sum_{j=1}^{q_i} N_{i,j} * \mathcal{H}_{\pi\text{-}up}(p_{i,j}, N_{i,j}, \gamma) \qquad (12)$$

It is easy to remark that in AIC and MDL, the accuracy of the BN (described by $LogP(\mathcal{B}, \mathcal{D})$) is computed independently of the complexity of the graph. Thus, even if it is clear that the number of examples used for evaluating the different lines of the tables decreases when $Dim(\mathcal{B})$ increases, it does not reflect all the possible situations (very homogeneous distributions of the data over the lines, or on the contrary very heterogeneous distributions, for instance). $POSS(\mathcal{B}, \mathcal{D})$ evaluates the amount of uncertainty on each conditional distribution and automatically gives a trade-off between uncertainty (related to the complexity of the graph) and the accuracy of the model.

In order to obtain the structure of the BN, a classical steepest hill climbing approach is used. However, we also propose a very simple incremental learning approach. For each new example, we apply the following process:

1. Update the score of each nodes
2. Update the score for each possible addition of an edge
3. If the addition of at least one edge increase the global score then add the edge that performs the best increase.

This approach can be done very efficiently for two reasons: *i)* each score function (*AIC*, *MDL*, *POSS*) can be decomposed into local score functions for each line of the tables, only the lines that correspond to the new example are updated, *ii)* the predicted score values for all the possible edge addition cans be stored in each line of the table and efficiently updated as in *i)*. For the sake of efficiency, no more than one edge can be added when considering a new example. This is reasonable since generally a BN contains far less nodes than the numbers of examples used for learning the structure and the tables. Since the *POSS* score considers the uncertainty of the conditional distributions locally, it appears to be suitable for this approach.

The only parameter of the algorithm is γ. It represents the strength of the constraint for uncertainty. This parameter can be automatically and effectively tuned very quickly by choosing the best value of $gamma$ for cross-validation in a small sub-sample of the training set (100 examples in the experiments).

5 Experimentation

In order to check the effectiveness of the proposed algorithms, we used 10 benchmarks from UCI[1] (numerical values are discretized). HAIC, HMDL and HPOSS

[1] `http://www.ics.uci.edu/~mlearn/MLRepository.html`

denote respectively a steepest hill climbing starting from an empty graph with the *AIC*, *MDL* and the *POSS* score functions. OAIC, OMDL and OPOSS corresponds to their online counterparts. The results in the following table corresponds to classification accuracy results for 10-cross validation. Departing from the normal use of these datasets, here all the variables of the dataset are regarded in turn as classes to be predicted from the remaining variables. We thus take into account the whole BN rather than only the nodes directly related to the variable that is usually taken as the class. Values in bold corresponds to statistically significant differences with the two other algorithms (Hill climbing and online algorithm are considered independently).

Data set	HAIC	HMDL	HPOSS	OAIC	OMDL	OPOSS
wine	77.5±2.2	77.3±2.4	78.2±2.4	75.4±2.1	73.7±3.5	**78.1±2.2**
diabetes	65.7±2.4	65.8±2.3	66.3±2.5	64.3±2.6	65.3±1.6	66.1±1.7
breast	79.6±1.9	79.9±1.7	80.0±1.6	78.9±2.4	79.4±2.0	79.7±1.7
vehicle	82.7±1.6	80.9±1.4	83.0±1.4	81.2±1.3	78.5±1.5	**82.3±1.1**
zoo	90.0±2.2	88.8±2.2	90.2±2.2	87.6±3.4	85.2±3.6	**90.6±1.7**
soybean	85.7±0.8	84.3±0.7	**88.6±0.5**	84.2±0.8	82.1±0.7	**87.6±0.4**
segment	74.1±1.0	70.9±1.0	**76.6±0.6**	73.7±0.9	59.1±1.4	73.8±0.7
glass	79.0±3.1	78.1±2.8	**83.7±2.2**	75.7±2.3	74.7±2.7	**82.0±3.4**
yeast	79.1±1.3	78.6±1.4	79.2±1.4	78.5±1.3	76.3±1.3	78.8±1.2
blocks	81.0±0.7	78.3±0.8	**83.4±0.7**	79.1±0.6	78.3±0.5	**81.3±0.4**

HPOSS statistically overcomes HAIC and HMDL on 4 of the 10 databases and is never overcome (statistically or not). When considering the online version, OAIC and OMDL algorithms obtain less good results than their hill climbing counterparts. On the opposite, OPOSS obtains similar results as HPOSS. OPOSS takes generally more time than HPOSS to learn a BN (which is easily understandable since it considers examples one by one) but the updating time is less than 1 ms in most cases and 39 ms in the worst case.

6 Conclusion

In this paper we have proposed an extension of the log-based entropy that takes into account the confidence intervals of the estimates of the frequencies with a limited amount of data, thanks to the use of a possibility-based representation of the family of probability distributions that agree with the data. We have shown that we can use this entropy directly as a score function to learn the structure of a Bayesian network. Experiments show that our algorithms perform very well again the classical information score functions and confirms the reliability and the efficiency of the online algorithm proposed. In the future, we shall compare more precisely our entropy measure with π-*up* entropy on a credal set. We also plan to investigate the learning of structures and conditional distributions when the data are incomplete. Besides, the tuning of the γ parameters in OPOSS could be made automatically during the updating process.

References

1. Abellàn, J., Moral, S.: Upper entropy of credal sets. applications to credal classification. International Journal of Approximate Reasoning 39, 235–255 (2005)
2. Agresti, A., Coull, B.: Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician 52(2), 119–126 (1998)
3. Dubois, D.: Possibility theory and statistical reasoning. Computational Statistics and Data Analysis 51, 47–69 (2006)
4. Dubois, D., Foulloy, L., Mauris, G., Prade, H.: Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliable Computing 10, 273–297 (2004)
5. Dubois, D., Prade, H.: When upper probabilities are possibility measures. Fuzzy Sets and Systems 49, 65–74 (1992)
6. Dubois, D., Prade, H., Sandri, S.: On possibility/probability transformations. In: Proceedings of Fourth IFSA Conference, pp. 103–112. Kluwer Academic Publ. (1993)
7. Heckerman, D., Chickering, D.M.: Learning bayesian networks: The combination of knowledge and statistical data. In: Machine Learning, pp. 20–197 (1995)
8. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
9. Serrurier, M., Prade, H.: An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. Int. J. Approx. Reasoning 54(7), 919–933 (2013)
10. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, 3–25 (1978)

On Fuzzy Equations with 2 × 2 Matrices

Martin Bacovský[*]

Faculty of Science Department of Mathematics,
University of Ostrava,
30. dubna 22, 701 03 Ostrava, Czech Republic
martin.bacovsky@osu.cz
http://www.osu.cz

Abstract. The question of probability of a system of fuzzy equations solvability in a max–t-norm fuzzy algebra for several t-norms and 2×2 matrices is considered. We derive that the probability of solving such a system is very low, namely $\frac{1}{10}$ for Gödel norm, $\frac{11}{60}$ for Łukasiewicz norm, $\frac{5}{36}$ for product norm and zero for drastic norm. These results are surprising compared to the case of a finite vector space, where the probability is one.

Keywords: t-norms, fuzzy relation equations, fuzzy algebras.

1 Introduction

Since the pioneer work [1] of fuzzy relation equations, many results on finding minimal and maximal elements and solvability of such systems have been developed (e.g. [2,3,4,5,6,7]). However, to the author's knowledge, there is no result describing something like probability of solving such a system. By probability is understood a situation, when one picks up uniformly randomly a matrix A together with a right-hand side b with coefficients from $[0; 1]$, then how often will the composed system $A \otimes x = b$ have a solution with respect to x?

This paper aims to make a first step in answering such kind of questions. To do this, we have restricted ourselves only to 2×2 matrices and four fuzzy algebras with t-norms minimum, Łukasiewicz, product and drastic. The derivation of conditional probabilities in these cases is presented. However, generalization to higher dimensions is the task of our future work.

After introduction, we continue by definitions of t-norms, fuzzy algebras and formulating rigorously the question of our concern. Then we answer the question in the case of finite vector spaces to have a classical result to which the results of fuzzy cases can be compared. The main section starts with results common for all t-norms. Then the respective cases are studied. The last section summarizes obtained results.

[*] This work was supported by grant SGS13/PřF/2014 of the University of Ostrava.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_12

2 Preliminaries

In this section, basic definitions, problem formulation and result for classical case are given.

Definition 1. *A mapping* $\mathrm{T} : [0;1] \times [0;1] \mapsto [0;1]$ *is called a* t-norm*, if the following conditions are satisfied for all* $a, b, c \in [0;1]$*:*

1. $a \,\mathrm{T}(b \,\mathrm{T}\, c) = (a \,\mathrm{T}\, b) \,\mathrm{T}\, c$ *(associativity),*
2. $a \,\mathrm{T}\, b = b \,\mathrm{T}\, a$ *(commutativity),*
3. *if* $a \leq b$*, then* $a \,\mathrm{T}\, c \leq b \,\mathrm{T}\, c$ *(monotonicity),*
4. $a \,\mathrm{T}\, 1 = a$ *(*1 *as neutral element).*

From all continuous t-norms (w.r.t. usual definition of continuity of a mapping), the following three play a prominent role, since every other continuous t-norm is their ordinal sum [9]:

$$a \,\mathrm{T}_G\, b := \min\{a, b\}, \ a \,\mathrm{T}_{\mathrm{L}}\, b := \max\{0, a + b - 1\}, \ a \,\mathrm{T}_{\Pi}\, b := ab.$$

The fourth one studied in this paper is the drastic norm:

$$a \,\mathrm{T}_D\, b = \begin{cases} 0 & \text{if } \max(a, b) < 1, \\ \min(a, b) & \text{otherwise,} \end{cases}$$

which is not continuous, but plays an important theoretical role, since it is the smallest possible t-norm.

Definition 2. *Given the unit interval* $[0;1]$ *together with a t-norm* T*, by* max–t-norm algebra *we understand an algebra* $\mathcal{A} = ([0;1], \max, \mathrm{T}, 0, 1)$ *of a type* $(2, 2, 0, 0)$*.*

Fuzzy algebras are usually viewed as complete residuated lattices (as e.g. in [8]). The chosen definition 2 is fully sufficient in what follows, since only the properties of t-norms are utilized.

We denote $\oplus := \max$ and $\otimes := \mathrm{T}$. The respective t-norm will be clear from the context. In notation, multiplication $\otimes$ is given precedence over addition $\oplus$, i.e., $a \ \otimes \ b \ \oplus \ c$ means the same as $(a \ \otimes \ b) \oplus \ c$.

Note, that $([0;1], \oplus, 0)$ and $([0;1], \otimes, 1)$ constitute commutative monoids and that T is distributive w.r.t. max: $a \otimes b \oplus a \otimes c = a \otimes (b \oplus c)$.

We extend the algebra $\mathcal{A}$ to a vector space-like structure $\mathcal{A}^n$ by formally replacing operations in standard matrix multiplication (matrix addition and multiplication by elements from a ring) by our operations $\oplus$ and $\otimes$. For example,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \otimes \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11} \otimes x_1 \oplus a_{12} \otimes x_2 \\ a_{21} \otimes x_1 \oplus a_{22} \otimes x_2 \end{pmatrix}.$$

Our question can be formulated as follows: *When randomly picking up a square matrix* $A \in \mathcal{A}^{n \times n}$ *and a column* $b \in \mathcal{A}^n$*, what is the* probability*, that the equation*

$$A \otimes x = b$$

is solvable in $x \in \mathcal{A}^n$*?*

By *probability* is meant the ratio of the volume of all solvable pairs (A, b) to the volume of all pairs. Answers for $n = 2$ and four t-norms are given in section 3, but before proceeding to fuzzy algebras, we answer our question for the case of finite vector spaces over a field $\mathbb{R}$:

- for a regular matrix every system of linear equations (SLE) is solvable,
- determinant of such a matrix is nonzero,
- non-regular matrices satisfy equation $\det A = 0$,
- the set of all non-regular matrices thus compose a set of a zero measure in $\mathbb{R}^{n^2}$,
- then the set of pairs (A, b) of SLE with A non-regular is of a zero measure in $\mathbb{R}^{n^2+n}$, too.

Thence, in this case, the answer is $P_{\mathbb{R}} = 1$.

3 max–t-norm Algebras

We restrict ourselves to 2×2 dimensional squares A

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \tag{1}$$

from $\mathcal{A}^{2\times 2}$. Notation

$$\begin{pmatrix} \Box_1 & \Box_2 \\ \Box_3 & \Box_4 \end{pmatrix}_{\Box_5}$$

represents the set of all such pairs (A, b) for which

$$a_{11}\Box_1 b_1,\ a_{12}\Box_2 b_1,\ a_{21}\Box_3 b_2,\ a_{22}\Box_4 b_2 \text{ and finally } b_1 \Box_5 b_2,$$

where $b = (b_1, b_2)^T$ is the right-hand side and $\Box_i \in \{<, >\}$ for $i = 1, 2, \dots 5$. We consider only strict inequalities, since coefficients are uniformly distributed and the probability of obtaining just one exact value is zero. When $\Box_5$ is omitted, there is no relation between b_1 and b_2.

The following cases are either unsolvable due to the monotonicity of t-norms, i.e., when $a, b \in [0; 1],\ a < b$ then $a \otimes x < b$ for all $x \in [0; 1]$, or their solvable parts have a zero measure:

$$\begin{array}{lllll}
\begin{pmatrix} < & < \\ < & < \end{pmatrix}, & & & & V = \frac{1}{9}, \\
\begin{pmatrix} > & < \\ > & < \end{pmatrix}, & \begin{pmatrix} < & > \\ < & > \end{pmatrix}, & & & V = \frac{1}{36}, \\
\begin{pmatrix} > & > \\ < & < \end{pmatrix}, & \begin{pmatrix} < & < \\ > & > \end{pmatrix}, & & & V = \frac{1}{9}, \\
\begin{pmatrix} > & < \\ < & < \end{pmatrix}_>, & \begin{pmatrix} < & > \\ < & < \end{pmatrix}_>, & \begin{pmatrix} < & < \\ > & < \end{pmatrix}_<, & \begin{pmatrix} < & < \\ < & > \end{pmatrix}_<, & V = \frac{1}{90}, \\
\begin{pmatrix} > & < \\ < & < \end{pmatrix}_<, & \begin{pmatrix} < & > \\ < & < \end{pmatrix}_<, & \begin{pmatrix} < & < \\ > & < \end{pmatrix}_>, & \begin{pmatrix} < & < \\ < & > \end{pmatrix}_>, & V = \frac{4}{90}.
\end{array}$$

V denotes the volume of such a set, e.g.

$$V\left(\begin{pmatrix} < & < \\ > & < \end{pmatrix}_<\right) = \int_0^1 db_2 \int_0^{b_2} db_1 \int_0^{b_1} da_{11} \int_0^{b_1} da_{12} \int_{b_2}^1 da_{21} \int_0^{b_2} da_{22} =$$
$$= \int_0^1 db_2 \int_0^{b_2} db_1 \; b_1^2 \; b_2 \; (1-b_2) = \frac{1}{90}.$$

We see, that the volume of unsolvable pairs must be strictly greater than

$$\frac{1}{9} + 2 \cdot \frac{1}{36} + 2 \cdot \frac{1}{9} + 4 \cdot \frac{1}{90} + 4 \cdot \frac{4}{90} = \frac{11}{18},$$

i.e., more than half of the whole volume of all pairs. On the other hand, the two following sets are always solvable for continuous t-norms (excluding thus also drastic norm):

$$\begin{pmatrix} > & < \\ < & > \end{pmatrix}, \begin{pmatrix} < & > \\ > & < \end{pmatrix}, \; V = \frac{1}{36}. \tag{2}$$

Thence, for continuous t-norms, the probability of system solvability with two equations and two variables is at least $\frac{1}{18}$. Solvability of other sets of pairs, namely

$$\begin{array}{ll} \mathrm{I}: & \begin{pmatrix} > & > \\ > & > \end{pmatrix}, \qquad\qquad\qquad\qquad\qquad\qquad V = \frac{1}{9}, \\ \mathrm{II}: & \begin{pmatrix} < & > \\ > & > \end{pmatrix}_>, \begin{pmatrix} > & < \\ > & > \end{pmatrix}_>, \begin{pmatrix} > & > \\ < & > \end{pmatrix}_<, \begin{pmatrix} > & > \\ > & < \end{pmatrix}_<, V = \frac{4}{90}, \\ \mathrm{III}: & \begin{pmatrix} < & > \\ > & > \end{pmatrix}_<, \begin{pmatrix} > & < \\ > & > \end{pmatrix}_<, \begin{pmatrix} > & > \\ < & > \end{pmatrix}_>, \begin{pmatrix} > & > \\ > & < \end{pmatrix}_>, V = \frac{1}{90}, \end{array}$$

substantially depends on the chosen t-norm, as is shown in next four subsections.

3.1 $\max -T_G$ Algebra

Sets of type I and II are not solvable in $\max -T_G$ algebra, since in the first case one of b_i's is strictly greater than other, w.l.o.g. let $b_1 > b_2$, then solving this one $x_1 = b_1$ leads to $\min(x_1, a_{21}) > b_2$ in the second equation. Similarly for $x_2 = b_1$. Case II differs only in that there is just one possibility for choosing x_i such that the equation with greater right-hand side is solved. Consider for example

$$(A, b) \in \begin{pmatrix} > & < \\ > & > \end{pmatrix}_>,$$

then clearly $x_1 = b_1$ in order to solve the first equation. But then $\min(a_{21}, x_1) > b_2$ and the second equation can not hold.

Systems from sets III are solvable; for instance for a pair (A, b) from

$$(A, b) \in \begin{pmatrix} > & > \\ < & > \end{pmatrix}_>$$

take $x = (b_1, b_2)$. The first equation is not corrupted, because $b_2 < b_1$, and neither is the second, because $a_{21} < b_2 < b_1$ and thus $\min(a_{21}, x_1) = a_{21} < b_2$.

The overall probability is $P_G = \dfrac{1}{18} + 4 \cdot \dfrac{1}{90} = \dfrac{1}{10}$, where $\frac{1}{18}$ comes from (2).

3.2 max $-T_{Ł}$ Algebra

In this algebra even systems from I and II may be solvable. Assume for example, that we would like to solve a system from I in variables a_{11} and a_{22}, i.e.,

$$x_1 = 1 + b_1 - a_{11},\ x_2 = 1 + b_2 - a_{22},$$

where $x_1, x_2 \le 1$ because $a_{ii} > b_i$ for $i = 1, 2$. Following conditions ensure, that neither of the equations will be corrupted

$$a_{21} + b_1 - a_{11} \le b_2,\ a_{12} + b_2 - a_{22} \le b_1. \tag{3}$$

It is now convenient to divide case I into two parts $b_1 > b_2$ and $b_2 > b_1$. For the first part, the conditions (3) can be rewritten to the form

$$a_{21} \le a_{11} + b_2 - b_1,\ a_{22} \ge a_{12} + b_2 - b_1$$

and $1 \ge a_{11} + b_2 - b_1 \ge b_2$ holds since $1 \ge a_{11} + \underbrace{b_2 - b_1}_{\le 0}$ and $\underbrace{a_{11} - b_1}_{\ge 0} + b_2 \ge b_2$, similarly for a_{22}. Then the volume of solvable part can be computed as

$$V\left(\begin{pmatrix} (>) & > \\ > & (>) \end{pmatrix}_{>}\right) = \\ = \int_0^1 db_1 \int_0^{b_1} db_2 \int_{b_1}^1 da_{11} \int_{b_2}^{a_{11}+b_2-b1} da_{21} \int_{b_1}^1 da_{12} \int_{a_{12}+b_2-b_1}^1 da_{22} = \\ = \frac{1}{80}.$$

Parenthesis indicate elements solving corresponding equation. Similarly can be dealt with other cases. The volumes of solvable parts of respective cases are

$$\text{I'}: \begin{pmatrix} (>) & > \\ > & (>) \end{pmatrix}_{>}, \begin{pmatrix} > & (>) \\ (>) & > \end{pmatrix}_{>}, \begin{pmatrix} (>) & > \\ > & (>) \end{pmatrix}_{<}, \begin{pmatrix} > & (>) \\ (>) & > \end{pmatrix}_{<}, V = \frac{1}{80},$$

and $V(\text{II}) = \frac{1}{80}$, $V(\text{III}) = \frac{1}{144}$. The probability is then

$$P_{Ł} = \frac{1}{18} + 4 \cdot \frac{1}{80} + 4 \cdot \frac{1}{80} + 4 \cdot \frac{1}{144} = \frac{11}{60}.$$

3.3 max $-T_{\Pi}$ Algebra

This algebra has very similar properties as the previous one, just replace subtraction by division and addition by multiplication, e.g. $a_{11} + b_2 - b_1$ now reads $a_{11}\dfrac{b_2}{b_1}$. Then the volumes of solvable parts are

$$V\left(\begin{pmatrix} (>) & > \\ > & (>) \end{pmatrix}_{>}\right) =$$
$$= \int_0^1 db_1 \int_0^{b_1} db_2 \int_{b_1}^1 da_{11} \int_{b_2}^{a_{11}\frac{b_2}{b1}} da_{21} \int_{b_1}^1 da_{12} \int_{a_{12}\frac{b_2}{b_1}}^1 da_{22} =$$
$$= \frac{1}{144},$$

$V(\mathrm{II}) = \dfrac{1}{180}$ and $V(\mathrm{III}) = \dfrac{1}{120}$. The probability in this case is

$$P_{\Pi} = \frac{1}{18} + 4 \cdot \frac{1}{144} + 4 \cdot \frac{1}{180} + 4 \cdot \frac{1}{120} = \frac{5}{36}.$$

3.4 max $-T_D$ Algebra

In this algebra the probability is zero, because in order to obtain a value b_i from a multiplication $A \otimes x$, there must be either a number 1 or b_i in the matrix A. However, the probability that this happens is zero.

4 Conclusions

In our contribution, we tried to make a first step in understanding what systems of linear equations (SLE) in classical vector spaces and fuzzy relation equations in fuzzy algebras have in common. We have shown computing with 2×2 matrices has completely different properties and from the point of view of solvability has nothing in common with SLE in vector spaces, since the probability of solvability of random pair (A, b) is very low in fuzzy algebras whilst in vector spaces it is certainly solvable. If we compare ordered t-norms with ordered obtained probabilities

$$\mathrm{T}_D \leq \mathrm{T}_{Ł} \leq \mathrm{T}_{\Pi} \leq \mathrm{T}_G,$$
$$P_D < P_G < P_{\Pi} < P_{Ł},$$
$$\frac{0}{180} < \frac{18}{180} < \frac{25}{180} < \frac{33}{180},$$

we can conjecture, that for two comparable continuous t-norms the less one (recall that $T_1 \leq T_2$ if for all $a, b \in [0;1]$ holds that $a\ T_1\ b \leq a\ T_2\ b$) will have greater probability of solvability.

In our next research, we will aim to answer the question for a general dimension.

References

1. Sanchez, E.: Resolution of Composite Fuzzy Relation Equations. Information and Control 30, 38–48 (1976)
2. Wagenknecht, M., Hartmann, K.: Fuzzy Modelling with Tolerances. Fuzzy Sets and Systems 20, 3 (1986)
3. Shieh, B.-S.: Solutions of fuzzy relation equations based on continuous t-norms. Information Sciences 177, 4208–4215 (2007)
4. Shieh, B.-S.: Deriving minimal solutions for fuzzy relation equations with max-product composition. Information Sciences 178, 3766–3774 (2008)
5. Czogala, E., Drewniak, J., Pedrycz, W.: Fuzzy relation equations in a finite set. Fuzzy Sets and Systems 7, 89–101 (1982)
6. Higashi, M., Klir, G.J.: Resolution of finite fuzzy relation equations. Fuzzy Sets and Systems 13, 65–82 (1984)
7. Gavalec, M.: Solvability and unique solvability of max-min fuzzy equations. Fuzzy Sets and Systems 124, 385–393 (2001)
8. Perfilieva, I., Novák, V.: System of fuzzy relation equations as a continuous model of IF-THEN rules. Information Sciences 177(16), 3218–3227 (2007)
9. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer, Dordrecht (2000)

Fuzzy and Set-Valued Stochastic Differential Equations with Solutions of Decreasing Fuzziness

Marek T. Malinowski

Faculty of Mathematics, Computer Science and Econometrics,
University of Zielona Góra,
ul. Prof. Z. Szafrana 4a, 65-516 Zielona Góra, Poland
M.Malinowski@wmie.uz.zgora.pl, malinowskimarek@poczta.fm

Abstract. We consider fuzzy stochastic differential equations in a new formulation. The equations that we examine possess solutions which are the fuzzy stochastic processes with trajectories of decreasing fuzziness in their consecutive values. This is a novelty. We give a theorem that guarantees existence and uniqueness of solutions. Some fuzzy stochastic differential equations are solved explicitly and some visualizations of simulations connected with their solutions are included. All the results can be applied immediately to set-valued stochastic differential equations.

Keywords: Fuzzy stochastic differential equation, fuzzy stochastic integral equation, set-valued stochastic differential equation.

1 Introduction

The dynamical systems operating in a random environment are often described by stochastic differential equations which involve stochastic integrals [16]. The system's states described by single values of a phase space are uncertain because of random factors and stochastic noises. On the other hand, in praxis there are often some questions if a concrete equation is a perfect one for considered phenomenon. This is caused by an imperfect knowledge of considered system and this uncertainty is not of stochastic type. Some measurements which are made to match an appropriate model are imprecise, vague or fuzzy. In particular, the parameters and the functional relationships in the systems may not be known precisely but only some sets of possible values may be determined or some linguistic variables may be used to describe them. The human perception lead to unclear and ambiguous descriptions of the observed systems. To model the systems with vagueness and fuzziness, the theory of fuzzy differential equations has been proposed [1,3,5,14]. This theory is developed independently and separately from the theory of the stochastic differential equations. These two different theories interlock in a notion of the fuzzy stochastic differential equation [2], [6]-[13], [15]. Such equations form a quite new topic of research and can be useful in mathematical models of the systems governed by some random forces and with fuzzy states. In [9]-[13] the author considers the fuzzy stochastic differential equations in a form which is a very natural generalization of the

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_13

crisp stochastic differential equations. Roughly speaking, in [9]-[13] the following equation is under considerations

$$x(t) = x_0 + \int_0^t f(s, x(s))ds + \int_0^t g(s, x(s))dB(s), \quad t \in [0, T],$$

where x_0 denotes a fuzzy random variable, f is a fuzzy-valued coefficient which is random, g is a crisp-valued random mapping, B is Brownian motion. In this paper, we reformulate the equation written above. Namely, we move the integrals from right to left and therefore consider equation

$$x(t) + (-1)\int_0^t f(s, x(s))ds + (-1)\int_0^t g(s, x(s))dB(s) = x_0, \quad t \in [0, T].$$

Such treatment would not have any significance if the equations were the crisp ones. However, in the fuzzy case the situation is completely different. The solutions to both the equations lay open some different geometrical properties. Namely, the diameter of the trajectory values of a solution to the first equation increases when time t increases, whereas it decreases for the second equation. This confirmes that the fuzzy modeling is much subtler and much richer than the crisp modeling. Although we consider the mappings x_0, f, g, x with values in the space of fuzzy sets of $\mathbb{R}^d$, the results of this paper can be rewritten in a framework of fuzzy and set-valued differential equations in M-type 2 Banach spaces. In [13] we considered the first equation written above in a set-up of fuzzy sets (and sets) of infinite dimensional M-type 2 Banach spaces.

2 Preliminaries

By the symbol $\mathfrak{K}(\mathbb{R}^d)$ we denote the family of all nonempty, compact and convex subsets of $\mathbb{R}^d$. In $\mathfrak{K}(\mathbb{R}^d)$ we consider the Hausdorff metric d_H which is defined by $d_H(A, B) := \max\big\{\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|\big\}$, where $\|\cdot\|$ denotes a norm in $\mathbb{R}^d$. Let $(\Omega, \mathcal{A}, P)$ be a complete probability space By $\mathcal{L}^p(\Omega, \mathcal{A}, P; \mathfrak{K}(\mathbb{R}^d))$ we denote the set of random sets F which are L^p-integrably bounded (see [9]-[11] for the details). A fuzzy set u in $\mathbb{R}^d$ is characterized by its membership function (denoted by u again) $u\colon \mathbb{R}^d \to [0, 1]$. By $\mathfrak{F}(\mathbb{R}^d)$ we denote a set of fuzzy sets $u\colon \mathbb{R}^d \to [0, 1]$ such that $[u]^\alpha \in \mathfrak{K}(\mathbb{R}^d)$ for every $\alpha \in [0, 1]$, where $[u]^\alpha := \{\, a \in \mathbb{R}^d : u(a) \geq \alpha \,\}$ for $\alpha \in (0, 1]$ and $[u]^0 := \mathrm{cl}\{\, a \in \mathbb{R}^d : u(a) > 0 \,\}$. The set $[u]^0$ is called the support of the fuzzy set u. By $\langle r\rangle$ we mean the characteristic function of the singleton $\{r\}$, $r \in \mathbb{R}^d$. Obviously, $\langle r\rangle \in \mathfrak{F}(\mathbb{R}^d)$. The addition $u + v$ and scalar multiplication ru in $\mathfrak{F}(\mathbb{R}^d)$ can be defined levelwise, i.e. $[u + v]^\alpha = [u]^\alpha + [v]^\alpha$, $[ru]^\alpha = r[u]^\alpha$, where $u, v \in \mathfrak{F}(\mathbb{R}^d)$, $r \in \mathbb{R}$ and $\alpha \in [0, 1]$. If for $u, v \in \mathfrak{F}(\mathbb{R}^d)$ there exists $w \in \mathfrak{F}(\mathbb{R}^d)$ such that $u = v + w$ then w is said to be the Hukuhara difference of u and v and we denote it by $u \ominus v$. In $\mathfrak{F}(\mathbb{R}^d)$ we consider the metric $d_\infty(u, v) := \sup_{\alpha\in[0,1]} d_H([u]^\alpha, [v]^\alpha)$. For $u \in \mathfrak{F}(\mathbb{R}^d)$ we denote an indicator of the fuzziness of u by $\mathrm{Fuzz}(u) := \mathrm{diam}([u]^0) = \sup\{\|a-b\| : a, b \in [u]^0\}$. If u is crisp then $\mathrm{Fuzz}(u) = 0$.

An $x\colon \Omega \to \mathfrak{F}(\mathbb{R}^d)$ is called a fuzzy random variable (see [17]), if $[x]^\alpha\colon \Omega \to \mathfrak{K}(\mathbb{R}^d)$ is a random set for all $\alpha \in [0,1]$. In [4] it is proved that $x\colon \Omega \to \mathfrak{F}(\mathbb{R}^d)$ is the fuzzy random variable if and only if $x\colon (\Omega, \mathcal{A}) \to (\mathfrak{F}(\mathbb{R}^d), \mathcal{B}_{d_S})$ is $\mathcal{A}|\mathcal{B}_{d_S}$-measurable, where d_S denotes the Skorohod metric in $\mathfrak{F}(\mathbb{R}^d)$ and $\mathcal{B}_{d_S}$ denotes the σ-algebra generated by the topology induced by d_S. A fuzzy random variable $x\colon \Omega \to \mathfrak{F}(\mathbb{R}^d)$ is said to be L^p-integrably bounded, $p \geq 1$, if $[x]^0$ belongs to $\mathcal{L}^p(\Omega, \mathcal{A}, P; \mathfrak{K}(\mathbb{R}^d))$. By $\mathcal{L}^p(\Omega, \mathcal{A}, P; \mathfrak{F}(\mathbb{R}^d))$ we denote the set of the all L^p-integrably bounded fuzzy random variables.

Denote $I := [0,T]$. We equip the probability space with a filtration $\{\mathcal{A}_t\}_{t\in I}$ satisfying the usual hypotheses. An $x\colon I \times \Omega \to \mathfrak{F}(\mathbb{R}^d)$ is called the fuzzy stochastic process, if for every $t \in I$ the mapping $x(t,\cdot)\colon \Omega \to \mathfrak{F}(\mathbb{R}^d)$ is a fuzzy random variable. It is d_∞-continuous, if almost all (with respect to the probability measure P) its trajectories, i.e. the mappings $x(\cdot,\omega)\colon I \to \mathfrak{F}(\mathbb{R}^d)$ are d_∞-continuous functions. A fuzzy stochastic process x is said to be nonanticipating, if for every $\alpha \in [0,1]$ the mapping $[x(\cdot,\cdot)]^\alpha$ is measurable with respect to the σ-algebra $\mathcal{N}$, which is defined as follows $\mathcal{N} := \{A \in \mathcal{B}(I) \otimes \mathcal{A} : A^t \in \mathcal{A}_t \text{ for every } t \in I\}$, where $A^t = \{\omega : (t,\omega) \in A\}$. Let $p \geq 1$ and $L^p(I \times \Omega, \mathcal{N}; \mathbb{R}^d)$ denote the set of all nonanticipating stochastic processes $h\colon I \times \Omega \to \mathbb{R}^d$ such that $\mathbb{E} \int_I \|h(s)\|^p ds < \infty$. A fuzzy stochastic process x is called L^p-integrably bounded ($p \geq 1$), if there exists a real-valued stochastic process $h \in L^p(I \times \Omega, \mathcal{N}; \mathbb{R})$ such that $d_\infty(x(t,\omega), \langle 0 \rangle) \leq h(t,\omega)$ for a.a. $(t,\omega) \in I \times \Omega$. By $\mathcal{L}^p(I \times \Omega, \mathcal{N}; \mathfrak{F}(\mathbb{R}^d))$ we denote the set of nonanticipating and L^p-integrably bounded fuzzy stochastic processes. For $\tau, t \in I$, $\tau < t$, and $x \in \mathcal{L}^1(I \times \Omega, \mathcal{N}; \mathfrak{F}(\mathbb{R}^d))$ we can define (see [9]-[11]) the fuzzy stochastic Lebesgue–Aumann integral $\Omega \ni \omega \mapsto \int_\tau^t x(s,\omega)ds \in \mathfrak{F}(\mathbb{R}^d)$ which is a fuzzy random variable.

For convenience, from now on, the phrase "with $P.1$" stands for "with probability one". Also we will write $x \overset{P.1}{=} y$ instead of $P\big(x = y\big) = 1$, where x, y are random elements. Also we will write $x(t) \overset{I\ P.1}{=} y(t)$ instead of $P\big(x(t) = y(t)\ \ \forall\, t \in I\big) = 1$, where x, y are the stochastic processes.

3 Fuzzy and Set-Valued Stochastic Differential Equations

We shall consider two kinds of fuzzy stochastic differential equations

$$x(t) \overset{I\ P.1}{=} x_0 + \int_0^t f(s, x(s))ds + \left\langle \int_0^t g(s, x(s))dB(s) \right\rangle, \tag{1}$$

$$x(t) + (-1)\int_0^t f(s, x(s))ds + \left\langle (-1)\int_0^t g(s, x(s))dB(s) \right\rangle \overset{I\ P.1}{=} x_0, \tag{2}$$

where $x_0\colon \Omega \to \mathfrak{F}(\mathbb{R}^d)$ is a fuzzy random variable, $f\colon I \times \Omega \times \mathfrak{F}(\mathbb{R}^d) \to \mathfrak{F}(\mathbb{R}^d)$ and $g\colon I \times \Omega \times \mathfrak{F}(\mathbb{R}^d) \to \mathbb{R}^d$. The first integral is the fuzzy stochastic Lebesgue–Aumann integral, whereas the second integral is the crisp stochastic Itô integral.

Let $\tilde{T} \in (0,T]$ and $\tilde{I} := [0,\tilde{T}]$. For the definition written below we assume that $\tilde{T} < T$.

Definition 1. *A fuzzy stochastic process* $x\colon \tilde{I} \times \Omega \to \mathfrak{F}(\mathbb{R}^d)$ *is said to be a local solution to Eq. (1), respectively to Eq. (2) if* $x \in \mathcal{L}^2(\tilde{I} \times \Omega, \mathcal{N}; \mathfrak{F}(\mathbb{R}^d))$, *x is* d_∞*-continuous, it satisfies (1) or satisfies (2), respectively, with* $\tilde{I}$ *instead of* I*. A local solution* $x\colon \tilde{I} \times \Omega \to \mathfrak{F}(\mathbb{R}^d)$ *to (1) (to (2), respectively) is said to be unique, if* $x(t) \overset{\tilde{I}\ P.1}{=} y(t)$*, where* $y\colon \tilde{I} \times \Omega \to \mathfrak{F}(\mathbb{R}^d)$ *is any local solution to (1) (to (2), respectively).*

If $\tilde{T} = T$ then a repetition of this definition gives the notions of global solutions and their uniqueness.

In [11] we proved that solutions to Eq. (1) possess trajectories with nondecreasing fuzziness in their values. Now, for the dual equation (2) we have the following assertion.

Theorem 1. *Assume that* $x\colon \tilde{I} \times \Omega \to \mathfrak{F}(\mathbb{R}^d)$ *is a (local or global) solution to Eq. (2). Then with* $P.1$ *the function* $\mathrm{Fuzz}(x(\cdot,\omega))\colon I \to \mathbb{R}$ *is nonincreasing.*

In fact, we can write even more. Namely under assumptions of this theorem we obtain that with $P.1$ for every $\alpha \in [0,1]$ the function $\mathrm{diam}([x(\cdot,\omega)]^\alpha)\colon I \to \mathbb{R}$ is nonincreasing.

To guarantee the existence and uniqueness of a solution to Eq. (2), we will assume that $f\colon I \times \Omega \times \mathfrak{F}(\mathbb{R}^d) \to \mathfrak{F}(\mathbb{R}^d)$, $g\colon I \times \Omega \times \mathfrak{F}(\mathbb{R}^d) \to \mathbb{R}^d$ satisfy:

(H1) the mapping $f\colon (I \times \Omega) \times \mathfrak{F}(\mathbb{R}^d) \to \mathfrak{F}(\mathbb{R}^d)$ is $\mathcal{N} \otimes \mathcal{B}_{d_S} | \mathcal{B}_{d_S}$-measurable and $g\colon (I \times \Omega) \times \mathfrak{F}(\mathbb{R}^d) \to \mathbb{R}^d$ is $\mathcal{N} \otimes \mathcal{B}_{d_S} | \mathcal{B}(\mathbb{R}^d)$-measurable,

(H2) there exists a constant $L > 0$ such that for $\lambda \times P$-a.a. (t,ω) and for every $u, v \in \mathfrak{F}(\mathbb{R}^d)$ it holds

$$\max\bigl\{d_\infty^2\bigl(f(t,\omega,u), f(t,\omega,v)\bigr), \|g(t,\omega,u) - g(t,\omega,v)\|^2\bigr\} \le L d_\infty^2(u,v),$$

(H3) there exists $C > 0$ such that for $\lambda \times P$-a.a. (t,ω)

$$\max\bigl\{d_\infty^2\bigl(f(t,\omega,\langle 0\rangle), \langle 0\rangle\bigr), \|g(t,\omega,\langle 0\rangle)\|^2\bigr\} \le C,$$

(H4) there exists $\tilde{T} \in (0,T]$ such that the sequence of the fuzzy stochastic processes $x_n\colon \tilde{I} \times \Omega \to \mathfrak{F}(\mathbb{R}^d)$ is well defined ($\tilde{I} = [0,\tilde{T}]$), where $x_0(t) \overset{\tilde{I}\ P.1}{:=} x_0$ and for $n = 1, 2, \dots$

$$x_n(t) \overset{\tilde{I}\ P.1}{:=} x_0 \ominus \Bigl[(-1)\int_0^t f(s, x_{n-1}(s))ds + \Bigl\langle (-1)\int_0^t g(s, x_{n-1}(s))dB(s)\Bigr\rangle\Bigr].$$

The condition (H3) is weaker than the following linear growth condition:

(H5) there exists $C > 0$ such that for $\lambda \times P$-a.a. (t,ω) and for every $u \in \mathfrak{F}(\mathbb{R}^d)$

$$\max\bigl\{d_\infty^2\bigl(f(t,\omega,u), \langle 0\rangle\bigr), \|g(t,\omega,u)\|^2\bigr\} \le C(1 + d_\infty^2(u, \langle 0\rangle)).$$

Theorem 2. *Let* $x_0 \in \mathcal{L}^2(\Omega, \mathcal{A}_0, P; \mathfrak{F}(\mathbb{R}^d))$*. Assume that (H1)-(H4) are satisfied. Then Eq. (2) has a unique, possibly local, solution defined on* $\tilde{I} \times \Omega$*.*

Corollary 1. *Let $x_0 \in \mathcal{L}^2(\Omega, \mathcal{A}_0, P; \mathfrak{F}(\mathbb{R}^d))$. Assume that (H1), (H2), (H4) and (H5) are satisfied. Then Eq. (2) has a unique, possibly local solution.*

A particular case of Eq. (2) is the following set-valued stochastic integral equation:

$$X(t) + (-1)\int_0^t F(s, X(s))ds + \left\{(-1)\int_0^t G(s, X(s))dB(s)\right\} \overset{I\ P.1}{=} X_0, \tag{3}$$

where $X_0 \colon \Omega \to \mathfrak{K}(\mathbb{R}^d)$ is a random set, $F \colon I \times \Omega \times \mathfrak{K}(\mathbb{R}^d) \to \mathfrak{K}(\mathbb{R}^d)$, $G \colon I \times \Omega \times \mathfrak{K}(\mathbb{R}^d) \to \mathbb{R}^d$ and the first integral is the set-valued stochastic Lebesgue–Aumann integral and the second integral is the crisp stochastic Itô integral.

Corollary 2. *Suppose that $X \colon \tilde{I} \times \Omega \to \mathfrak{K}(\mathbb{R}^d)$ is a (local or global) solution to Eq. (3). Then with $P.1$ the function $\operatorname{diam}(X(\cdot, \omega)) \colon I \to \mathbb{R}$ is nonincreasing.*

Reformulating the conditions (H1)-(H4) into the following conditions:

(A1) the mapping $F \colon (I \times \Omega) \times \mathfrak{K}(\mathbb{R}^d) \to \mathfrak{K}(\mathbb{R}^d)$ is $\mathcal{N} \otimes \mathcal{B}_{d_H} | \mathcal{B}_{d_H}$-measurable and $G \colon (I \times \Omega) \times \mathfrak{K}(\mathbb{R}^d) \to \mathbb{R}^d$ is $\mathcal{N} \otimes \mathcal{B}_{d_H} | \mathcal{B}(\mathbb{R}^d)$-measurable,

(A2) there exists a constant $L > 0$ such that for $\lambda \times P$-a.a. (t, ω) and for every $A, B \in \mathfrak{K}(\mathbb{R}^d)$ it holds

$$\max\left\{d_H^2\big(F(t,\omega,A), F(t,\omega,B)\big), \|G(t,\omega,A) - G(t,\omega,B)\|^2\right\} \le L d_H^2(A,B),$$

(A3) there exists $C > 0$ such that for $\lambda \times P$-a.a. (t, ω)

$$\max\left\{d_H^2\big(F(t,\omega,\{0\}), \{0\}\big), \|G(t,\omega,\{0\})\|^2\right\} \le C,$$

(A4) there exists $\tilde{T} \in (0, T]$ such that the sequence of the set-valued stochastic processes $X_n \colon \tilde{I} \times \Omega \to \mathfrak{K}(\mathbb{R}^d)$ is well defined ($\tilde{I} = [0, \tilde{T}]$), where $X_0(t) \overset{\tilde{I}\ P.1}{:=} X_0$ and for $n = 1, 2, \ldots$

$$X_n(t) \overset{\tilde{I}\ P.1}{:=} X_0 \ominus \Big[(-1)\int_0^t F(s, X_{n-1}(s))ds + \Big\{(-1)\int_0^t G(s, X_{n-1}(s))dB(s)\Big\}\Big].$$

we obtain immediately the following assertion.

Corollary 3. *Let $X_0 \in \mathcal{L}^2(\Omega, \mathcal{A}_0, P; \mathfrak{K}(\mathbb{R}^d))$. Suppose that (A1)-(A4) are satisfied. Then Eq. (3) has a unique, possibly local solution defined on $\tilde{I} \times \Omega$.*

4 Examples and Numerical Simulations

It is difficult to find explicit formulae for solutions to crisp stochastic differential equations. Fuzzy stochastic equations inherit this kind of a heavy task. However, some simple equations possess solutions that can be written in a closed, explicit form. In what follows we consider some examples of fuzzy stochastic differential equations with values in $\mathfrak{F}(\mathbb{R})$. To analyze an essential difference between Eqs (1) and (2), we consider the following fuzzy stochastic differential equations

$$x(t) \;{}^{I}\overset{P.1}{=}\; x_0 + \int_0^t \xi(s)x(s)ds + \Big\langle \int_0^t \theta(s)dB(s) \Big\rangle \tag{4}$$

and

$$x(t) + (-1)\int_0^t \xi(s)x(s)ds + \Big\langle (-1)\int_0^t \theta(s)dB(s) \Big\rangle \;{}^{I}\overset{P.1}{=}\; x_0, \tag{5}$$

where $\xi\colon I \to \mathbb{R}$ and $\theta\colon I \to (0,\infty)$ are measurable and bounded, $x_0\colon \Omega \to \mathfrak{F}(\mathbb{R})$ is a fuzzy random variable. These equations can be used as some models of population dynamics. They are some extensions of the well-known Malthus model in population modeling. The solutions to these equations are different when $\xi(t) \geq 0$ and $\xi(t) \leq 0$. Therefore we investigate these two cases separately. Notice that in the crisp case there is no need to such separate examinations. Moreover, in the crisp case Eqs (4) and (5) coincide. The fuzzy environment that we study results in a much richer examinations concerning Eqs (4) and (5).

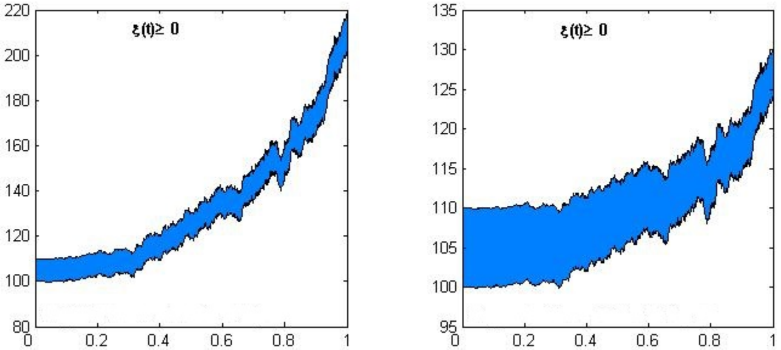


Fig. 1. The graphs of $[x(\cdot,\omega^*)]^0$ for the solutions x to (4) and (5) with $\xi(t) \geq 0$

Assume that $\xi(t) \geq 0$ for $t \in I$. After some calculations we obtain that the solution $x\colon I \times \Omega \to \mathfrak{F}(\mathbb{R})$ to (4) with non-negative ξ is as follows:

$$x(t) \;{}^{I}\overset{P.1}{=}\; \exp\left\{\int_0^t \xi(s)ds\right\} \cdot x_0 + \Big\langle \int_0^t \theta(s)\exp\left\{\int_s^t \xi(\tau)d\tau\right\} dB(s) \Big\rangle$$

and $\mathrm{Fuzz}(x(t)) = \mathrm{Fuzz}(x_0)\exp\left\{\int_0^t \xi(s)ds\right\}$ is nondecreasing as t increases. The solution $x\colon \tilde{I} \times \Omega \to \mathfrak{F}(\mathbb{R})$ to (5) with non-negative ξ is of the form

$$x(t) \ {}^{\tilde{I}}\!\overset{P.1}{=} \left[\cosh\left\{\int_0^t \xi(s)ds\right\} \cdot x_0 \ominus \left(-\sinh\left\{\int_0^t \xi(s)ds\right\} \cdot x_0\right)\right] \\ + \left\langle \int_0^t \theta(s) \exp\left\{\int_s^t \xi(\tau)d\tau\right\} dB(s)\right\rangle$$

and $\mathrm{Fuzz}(x(t)) = \mathrm{Fuzz}(x_0)\exp\left\{-\int_0^t \xi(s)ds\right\}$ is nonincreasing as t increases.

In the sequel we shall simulate numerically some trajectories of the solution supports $[x(\cdot,\omega)]^0$ corresponding to the equations (4) and (5) with non-negative ξ. In this way, in Fig. 1, we illustrate a behavior of the fuzziness of the solutions values. One can see the monotonicity of fuzziness of the trajectory values $\mathrm{Fuzz}(x(\cdot,\omega))$. Let us put the following data in (4) and (5): $I = [0,1]$, $\xi(t) = t$ and $\theta(t) = t$ for $t \in I$. Assume that for an $\omega^* \in \Omega$ the realization of the support of the initial value is as follows $[x_0(\omega^*)]^0 = [100, 110]$. For ω^* we simulate one trajectory of the solution support to both the equations. A corresponding illustration is drawn in Fig. 1.

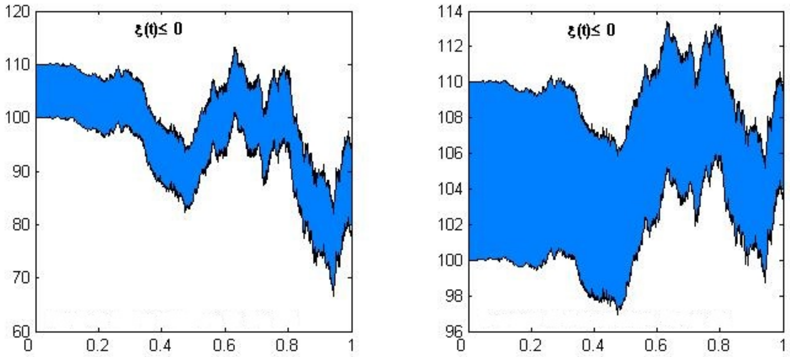


Fig. 2. The graphs of $[x(\cdot,\omega^*)]^0$ for the solutions x to (4) and (5) with $\xi(t) \le 0$

Now we assume that $\xi(t) \le 0$ for $t \in I$. Then we obtain that the solution $x\colon I \times \Omega \to \mathfrak{F}(\mathbb{R})$ to (4) with non-positive ξ reads:

$$x(t) \ {}^{I}\!\overset{P.1}{=} \cosh\left\{\int_0^t \xi(s)ds\right\} \cdot x_0 + \sinh\left\{\int_0^t \xi(s)ds\right\} \cdot x_0 \\ + \left\langle \int_0^t \theta(s) \exp\left\{\int_s^t \xi(\tau)d\tau\right\} dB(s)\right\rangle$$

and $\mathrm{Fuzz}(x(t)) = \mathrm{Fuzz}(x_0)\exp\left\{-\int_0^t \xi(s)ds\right\}$ is nondecreasing. The solution $x\colon \tilde{I} \times \Omega \to \mathfrak{F}(\mathbb{R})$ to (5) with non-positive ξ is of the form

$$x(t) \ {}^{\tilde{I}}\!\overset{P.1}{=} \exp\left\{\int_0^t \xi(s)ds\right\} \cdot x_0 + \left\langle \int_0^t \theta(s) \exp\left\{\int_s^t \xi(\tau)d\tau\right\} dB(s)\right\rangle$$

and $\mathrm{Fuzz}(x(t)) = \mathrm{Fuzz}(x_0)\exp\left\{\int_0^t \xi(s)ds\right\}$ is nonincreasing.

For a visualization of the solution support in the models (4) and (5) we set the data $I = [0, 1]$, $\xi(t) = -t$ and $\theta(t) = t$ for $t \in I$. Similarly like above we assume that for an $\omega^* \in \Omega$ the realization of the support of the initial value is $[x_0(\omega^*)]^0 = [100, 110]$. As a result of a simulation we obtain the trajectories $[x(\cdot, \omega^*)]^0$ corresponding to the solutions to (4) and (5). They are presented in Fig. 2. The illustrations Fig. 1 and Fig. 2 reflect the fact that the trajectories of the solution to Eq. (1) have nondecreasing fuzziness in their consecutive values, whereas the trajectories of the solution to Eq. (2) have nonincreasing fuzziness in their consecutive values.

References

1. Barros, L.C., Bassanezi, R.C., Tonelli, P.A.: Fuzzy modelling in population dynamics. Ecological Modelling 128, 27–33 (2000)
2. Feng, Y.: Fuzzy stochastic differential systems. Fuzzy Sets Syst. 115, 351–363 (2000)
3. Kaleva, O.: Fuzzy differential equations. Fuzzy Sets Syst. 24, 301–317 (1987)
4. Kim, Y.K.: Measurability for fuzzy valued functions. Fuzzy Sets Syst. 129, 105–109 (2002)
5. Lakshmikantham, V., Mohapatra, R.N.: Theory of Fuzzy Differential Equations and Inclusions. Taylor & Francies, London (2003)
6. Malinowski, M.T.: On random fuzzy differential equations. Fuzzy Sets Syst. 160, 3152–3165 (2009)
7. Malinowski, M.T.: Existence theorems for solutions to random fuzzy differential equations. Nonlinear Anal. TMA 73, 1515–1532 (2010)
8. Malinowski, M.T.: Random fuzzy differential equations under generalized Lipschitz condition. Nonlinear Anal. Real World Appl. 13, 860–881 (2012)
9. Malinowski, M.T.: Strong solutions to stochastic fuzzy differential equations of Itô type. Math. Comput. Modelling 55, 918–928 (2012)
10. Malinowski, M.T.: Itô type stochastic fuzzy differential equations with delay. Systems Control Lett. 61, 692–701 (2012)
11. Malinowski, M.T.: Some properties of strong solutions to stochastic fuzzy differential equations. Inform. Sci. 252, 62–80 (2013)
12. Malinowski, M.T.: Modeling with stochastic fuzzy differential equations. In: Chakraverty, S. (ed.) Mathematics of Uncertainty Modeling in the Analysis of Engineering and Science Problems, pp. 150–172. IGI Global, Hershey Pennsylvania (2014)
13. Malinowski, M.T.: Set-valued and fuzzy stochastic differential equations in M-type 2 Banach spaces. Tohoku Math. J. (to appear, 2014)
14. Nieto, J.: The Cauchy problem for continuous fuzzy differential equations. Fuzzy Sets Syst. 102, 259–262 (1999)
15. Ogura, Y.: On stochastic differential equations with fuzzy set coefficients. In: Dubois, D., et al. (eds.) Soft Methods for Handling Variability and Imprecision, ASC, vol. 48, pp. 263–270. Springer, Berlin (2008)
16. Øksendal, B.: Stochastic Differential Equations: An Introduction with Applications. Springer, Berlin (2003)
17. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 91, 552–558 (1983)

Part II
Soft Methods in Statistics

An Approximation to the Small Sample Distribution of the Trimmed Mean for Gaussian Mixture Models

Alfonso 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
agar-per@ccia.uned.es

Abstract. The α-trimmed mean, a statistic commonly used in robustness studies, has an intractable small sample distribution. For this reason, an asymptotic normal distribution or a Student t distribution are commonly used as approximations when the sample size is small. In this article we obtain an approximation for the small sample distribution of the α-trimmed mean, based on the von Mises expansion of a functional, which is valid for the case in which the observations come from a Gaussian Mixture Model.

Keywords: Robustness, α-trimmed mean, von Mises expansion.

1 Introduction

The α-trimmed mean is a very popular robust statistic used for location problems. If we trim the $100 \cdot \alpha\%$ of the smallest and the $100 \cdot \alpha\%$ of the largest ordered sample data $X_{(i)}$, the symmetrically α-trimmed mean is defined by

$$\overline{X}_\alpha = \frac{1}{n-2k}\left(X_{(k+1)} + ... + X_{(n-k)}\right)$$

where $k = [n\alpha]$ if $[\,.\,]$ stands for the integer part.

Its exact distribution is intractable (see for instance [13] pp. 31). Its large-sample approximation is asymptotically normal under some conditions although more complicated than for other L-estimates; see for instance [12] pp. 361, [13] pp. 31, [1] or [15].

When the sample size is small and the data are normally distributed, a Student's t distribution is used as an approximation for the standardized trimmed mean; see for instance [14] pp. 105 or pp. 156-157, or [16]. In fact, if it is

$$W_i = \begin{cases} X_{(k+1)} \ , & X_i \leq X_{(k+1)} \\ X_i \ , & X_{(k+1)} < X_i \leq X_{(n-k)} \\ X_{(n-k)} \ , & X_i \geq X_{(n-k)} \end{cases}$$

* The author is very grateful to two anonymous referees for their comments. This work was partially supported by the Grant MTM2012-33740.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_14

and $\overline{x}_\alpha^W$ is the α-Winsorized mean

$$\overline{x}_\alpha^W = \frac{1}{n}\sum_{i=1}^{n} W_i$$

being also the α-Winsorized quasi-variance

$$S_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(W_i - \overline{x}_\alpha^W\right)^2$$

then it is

$$\frac{\overline{X}_\alpha - \mu_\alpha}{\sqrt{\widehat{V}(\overline{X}_\alpha)}} = \frac{(1-2\alpha)\sqrt{n}\,(\overline{X}_\alpha - \mu_\alpha)}{S_W} \approx t_{n-2k-1}$$

where

$$\mu_\alpha = \frac{1}{1-2\,\alpha}\int_\alpha^{1-\alpha} F^{-1}(p)\,dp = \frac{1}{1-2\,\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y\,dF(y)$$

is the functional associated with the trimmed mean $\overline{X}_\alpha$.

Nevertheless, if the data are supposed to come from a normal distribution (i.e., no contamination is assumed) the trimmed mean is not really needed.

There are some Edgeworth expansions used as approximations, [10], but it is well known that these approximations are accurate only in the center of the distribution and not in the tails where they can even be negative.

The only accurate approximations for the distribution of $\overline{X}_\alpha$, when the sample size is small and the distribution not normal, are the saddlepoint approximations given in [11] or [2], although these are almost impossible to apply and the elements involved in them, difficult to interpret.

In some articles, [3], [4], [5], [6], [7], [8] and [9], a linear approximation, based on a von Mises expansion plus an iterative procedure, was used to obtain accurate approximations of some classical statistics when the underlying model is close to the normal distribution. In these articles a saddlepoint approximation was used in the computation of the Tail Area Influence Function (TAIF) that appears in the von Mises expansion. But, in two recent articles, [8] and [9], a new expression to compute exactly the TAIF was obtained, formula that can be used in the von Mises expansion instead of the saddlepoint approximation.

We shall use the von Mises expansion in combination with the exact expression of the TAIF, to obtain an accurate approximation to the small sample distribution of the trimmed mean when the underlying model is close to the normal.

2 Definitions and Computations

Although the random variables X_i in the sample $(X_1, ..., X_n)$ are independent and identically distributed (iid), in this section we shall consider statistics (e.g., the trimmed mean) for which it could be $T_n(X_1+c, X_2, ..., X_n) \neq T_n(X_1, X_2+c, ..., X_n)$ for a constant c. For this reason, in the following we shall consider statistics $T_n(X_1, ..., X_n)$ based on independent but not necessarily identically distributed univariate random variables X_i, being $X_i \equiv G_i$, $i = 1, ..., n$ ($X \equiv H$ stands for "X is distributed as H"), statistics that, in the case of a hypothesis testing problem, will reject the null hypothesis (usually about a parameter $\theta \in \Theta$) for large values of T_n, although the results can easily be extended to other situations.

Under very general conditions (Section 2 in [17]) we can use the first-order von Mises expansion (see Corollary 2 in [9]) to compute the tail probability functional under a model $\mathbf{F} = (F_1, ..., F_n)$ as

$$P_{\mathbf{F}}\{T_n(X_1, X_2, ..., X_n) > t\} = P_{F_1,...,F_n}\{T_n(X_1, X_2, ..., X_n) > t\} =$$

$$= P_{\mathbf{G}}\{T_n(X_1, X_2, ..., X_n) > t\} + \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i\ (x; t; T_n, \mathbf{G})\ dF_i(x) + Rem$$

where TAIF_i is the i-th Partial Tail Area Influence Function of T_n at $\mathbf{G} = (G_1, ..., G_n)$ with relation to G_i, $i = 1, ..., n$, defined in [9] by

$$\mathrm{TAIF}_i(x; t; T_n, \mathbf{G}) = \left. \frac{\partial}{\partial \epsilon} P_{G_i^{\epsilon,x}}\{T_n(X_1, ..., X_n) > t\} \right|_{\epsilon=0}$$

in those $x \in \mathcal{X}$ where the right hand side exists, being $G_i^{\epsilon,x} = (1-\epsilon)G_i + \epsilon\,\delta_x$, $i = 1, ..., n$, and δ_x the probability measure which assigns mass 1 at the point $x \in \mathcal{X} \subset \mathbb{R}$.

In the computation of the TAIF_i only G_i is contaminated; the other distributions remain fixed, $i = 1, ..., n$.

Here we assume this situation and also that the X_i's are univariate although an extension to multivariate case would be straightforward (see [9]).

The remainder term

$$Rem = \frac{1}{2} \int \int T^{(2)}_{\mathbf{G_F}}(x_1, x_2)\, d[\mathbf{F}(x_1) - \mathbf{G}(x_1)]\, d[\mathbf{F}(x_2) - \mathbf{G}(x_2)]$$

is small if distributions $\mathbf{F}$ and $\mathbf{G}$ are close. ($T^{(2)}_{\mathbf{G_F}}$ is the *second derivative* of the tail probability functional at the mixture distribution $\mathbf{G_F} = (1-\lambda)\mathbf{G} + \lambda\mathbf{F}$, for some $\lambda \in [0, 1]$.)

Hence, if $\mathbf{F}$ and $\mathbf{G}$ are close enough, we can write, using the exact expression for the TAIF_i obtained in [9]

$$P_{\mathbf{F}}\{T_n(X_1, X_2, ..., X_n) > t\} \simeq P_{\mathbf{G}}\{T_n(X_1, X_2, ..., X_n) > t\} + \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i\ (x; t; T_n, \mathbf{G})\, dF_i(x) \tag{1}$$

$$= (1-n)P_{\mathbf{G}}\{T_n(X_1, X_2, ..., X_n) > t\} + \int_{\mathcal{X}} P_{G_2,...,G_n}\{T_n(x, X_2, ..., X_n) > t\}\, dF_1(x) +$$

$$+\int_{\mathcal{X}} P_{G_1,G_3,...,G_n}\{T_n(X_1,x,...,X_n)>t\}\,dF_2(x)+\cdots$$

$$+\int_{\mathcal{X}} P_{G_1,...G_{n-1}}\{T_n(X_1,...,X_{n-1},x)>t\}\,dF_n(x) \tag{2}$$

that allows an approximation of the tail probability $P_{\mathbf{F}}\{T_n > t\}$ under models $(F_1,...,F_n)$, knowing the value of this tail probability under near models $(G_1,...,G_n)$.

In order to value the influence of outliers, we shall consider as model $\mathbf{F} = (1-\epsilon)\mathbf{G} + \epsilon\mathbf{G}_s$ where $\mathbf{G}_s$ is a shift version of $\mathbf{G}$ and $\epsilon \in [0, 0.5]$ a parameter which measures the contamination.

Namely, if $\mathbf{G}$ are location families with a common location parameter θ_0, we shall suppose that $\mathbf{G}_s$ have a common location parameter $\theta > \theta_0$.

In this case, we shall have, for instance, in the last integral (2), if $t = t_n$ is a possible value of T_n and φ the random function (test or critical function in a hypothesis testing problem)

$$\varphi(x_1,...,x_n) = \begin{cases} 1 & if \quad T_n(x_1,x_2,...,x_n) > t_n \\ 0 & if \quad T_n(x_1,x_2,...,x_n) \le t_n \end{cases}$$

that

$$\int_{\mathcal{X}} P_{G_{1;\theta_0},...G_{n-1;\theta_0}}\{T_n(X_1,...,X_{n-1},x)>t_n\}\,dF_n(x)$$

$$=(1-\epsilon)\int_{\mathcal{X}} P_{G_{1;\theta_0},...G_{n-1;\theta_0}}\{T_n(X_1,...,X_{n-1},x)>t_n\}\,dG_{n;\theta_0}(x)$$

$$+\epsilon\int_{\mathcal{X}} P_{G_{1;\theta_0},...G_{n-1;\theta_0}}\{T_n(X_1,...,X_{n-1},x)>t_n\}\,dG_{n;\theta}(x)$$

$$=(1-\epsilon)\int_{\mathcal{X}}\left[\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}\varphi(x_1,...,x_{n-1},x)\,dG_{1;\theta_0}(x_1)\cdots dG_{n-1;\theta_0}(x_{n-1})\right]\,dG_{n;\theta_0}(x)$$

$$+\epsilon\int_{\mathcal{X}}\left[\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}\varphi(x_1,...,x_{n-1},x)\,dG_{1;\theta_0}(x_1)\cdots dG_{n-1;\theta_0}(x_{n-1})\right]\,dG_{n;\theta}(x)$$

$$=(1-\epsilon)P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,...,X_n)>t_n\}+\epsilon P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,...,X_n+(\theta-\theta_0))>t_n\}$$

moving the shift parameter in the last integral with a simple change of variable. Hence, if $\mathbf{F} = (1-\epsilon)\mathbf{G}_{\theta_0} + \epsilon\mathbf{G}_\theta$

$$P_{\mathbf{F}}\{T_n(X_1,X_2,...,X_n)>t_n\} \simeq (1-\epsilon\, n)P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n\}+$$
$$+\epsilon\left(P_{\mathbf{G}_{\theta_0}}\{T_n(X_1+(\theta-\theta_0),X_2,...,X_n)>t_n\}+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2+(\theta-\theta_0),...,X_n)>t_n\}\right.$$

$$+\cdots+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n+(\theta-\theta_0))>t_n\}\big)$$

$$=(1-\epsilon\, n)P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,x_2,...,x_n)\}+\epsilon$$

$$\big(P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1-(\theta-\theta_0),x_2,...,x_n)\}$$

$$+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,x_2-(\theta-\theta_0),...,x_n)\}$$

$$+\cdots+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,x_2,...,x_n-(\theta-\theta_0))\}\big)\,.$$

3 Iterative Procedure

The previous approximation is accurate if $\mathbf{F}=(1-\epsilon)\mathbf{G}_{\theta_0}+\epsilon\mathbf{G}_\theta$ is close to $\mathbf{G}_{\theta_0}$, i.e., if ϵ is small and/or θ is close to θ_0. Nevertheless, in some situations, ϵ is not small or θ is far from θ_0. In these cases we can use an alternative iterative procedure considering intermediate distributions between $\mathbf{G}_{\theta_0}$ and $\mathbf{F}=\mathbf{G}_{(1-\epsilon)\theta_0+\epsilon\theta}$; namely, distributions $\mathbf{F}_j=(F_{1;\theta_j},...,F_{n;\theta_j})=(F_{1j},...,F_{nj})=\mathbf{G}_{\theta_0+(\theta-\theta_0)\epsilon j/(k+1)}$, $j=1,...,k+1$, where $\mathbf{F}_0=\mathbf{G}_{\theta_0}=(G_{1;\theta_0},...,G_{n;\theta_0})$ and $\mathbf{F}_{k+1}=\mathbf{F}=\mathbf{G}_{\theta_0+(\theta-\theta_0)\epsilon}$. With k iterations, equation (1) now becomes

$$P_{\mathbf{F}}\{T_n(X_1,X_2,...,X_n)>t_n\}\simeq P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n\}+$$

$$+\sum_{j=1}^{k+1}\int_{\mathcal{X}}\sum_{i=1}^{n}\mathrm{TAIF}_i\left(x;t_n;T_n,\mathbf{F}_{j-1}\right)\,dF_{ij}(x)$$

Moreover, since

$$\mathrm{TAIF}_i\left(x;t_n;T_n,\mathbf{F}_{j-1}\right)=$$

$$=P_{(F_{1,j-1},...,F_{i-1,j-1},F_{i+1,j-1},...,F_{n,j-1})}\left\{T_n(X_1,...,X_{i-1},x,X_{i+1},...,X_n)>t_n\right\}$$

$$-P_{\mathbf{F}_{j-1}}\left\{T_n(X_1,...,X_n)>t_n\right\}$$

if we consider again a location family as underlying distribution, i.e., that $\mathbf{F}_j=(F_{1;\theta_j},...,F_{n;\theta_j})=(F_{1j},...,F_{nj})$ is a location family with location parameter $\theta_j=\theta_0+\epsilon j(\theta-\theta_0)/(k+1)$ and the random function φ, we can move again the shift parameter in the distribution to the random variable with a change of variable, obtaining

$$P_{\mathbf{F}}\{T_n(X_1,X_2,...,X_n)>t_n\}\simeq P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n\}+$$

$$+\sum_{j=1}^{k+1}[P_{\mathbf{G}_0}\left\{T_n(X_1+c_{2j},X_2+c_{1j},...,X_n+c_{1j})>t_n\right\}$$

$$+P_{\mathbf{G}_0}\left\{T_n(X_1+c_{1j},X_2+c_{2j},X_3+c_{1j},...,X_n+c_{1j})>t_n\right\}$$

$$+...+P_{\mathbf{G}_0}\left\{T_n(X_1+c_{1j},...,X_{n-1}+c_{1j},X_n+c_{2j})>t_n\right\}$$

$$-nP_{\mathbf{G}_0}\left\{T_n(X_1+c_{1j},...,X_n+c_{1j})>t_n\right\}]$$

where $c_{1j}=\epsilon(j-1)(\theta-\theta_0)/(k+1)$ and $c_{2j}=\epsilon j(\theta-\theta_0)/(k+1)$. Hence,

$$P_{\mathbf{F}}\{T_n(X_1,X_2,...,X_n)>t_n\}\simeq P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,...,x_n)\}$$

$$+\sum_{j=1}^{k+1}[P_{\mathbf{G}_0}\left\{T_n(X_1,X_2,...,X_n)>t_n(x_1-c_{2j},x_2-c_{1j},...,x_n-c_{1j})\right\}$$

$$+P_{\mathbf{G}_0}\left\{T_n(X_1,X_2,...,X_n)>t_n(X_1-c_{1j},x_2-c_{2j},...,x_n-c_{1j})\right\}$$

$$+...+P_{\mathbf{G}_0}\left\{T_n(X_1,X_2,...,X_n)>t_n(x_1-c_{1j},...,,x_n-c_{2j})\right\}$$

$$-nP_{\mathbf{G}_0}\left\{T_n(X_1,X_2,...,X_n)>t_n(x_1-c_{1j},...,x_n-c_{1j})\right\}].$$

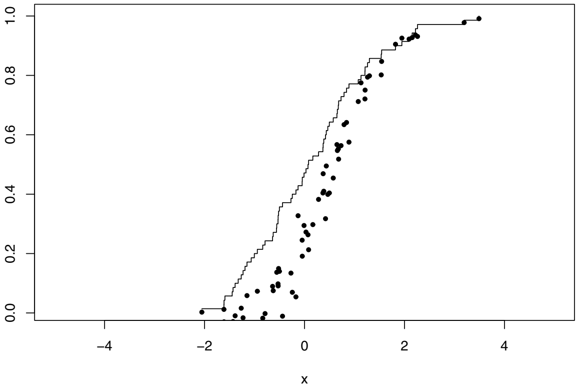


Fig. 1. Simulated (solid line) and von Mises approximation given by (3) (dotted) distributions of T_n with a $N((1-\epsilon)\theta_0+\epsilon\theta,1)$ model

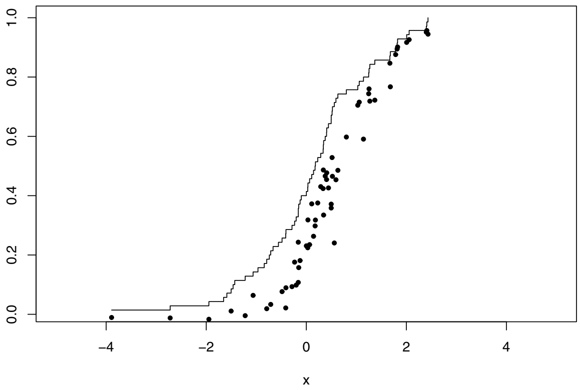


Fig. 2. Simulated (solid line) and von Mises approximation given by (3) (dotted) distributions of T_n with a $(1-\epsilon)N(\theta_0,1)+\epsilon N(\theta,1)$ model

If we considere now the standardized trimmed mean

$$T_n = \frac{\overline{X}_\alpha - \mu_\alpha}{\sqrt{\widehat{V}(\overline{X}_\alpha)}} = \frac{(1-2\alpha)\sqrt{n}\,(\overline{X}_\alpha - \mu_\alpha)}{S_W} \approx t_{n-2k-1}$$

and, as $\mathbf{G}_0$, standard normal distributions, for which we know that $T_n \approx t_{n-2k-1}$ we have

$$P_{G_{(1-\epsilon)\theta_0+\epsilon\theta}}\{T_n(X_1,X_2,...,X_n)>t_n\} \simeq P\{W>t_n(x_1,...,x_n)\}$$
$$+\sum_{j=1}^{k+1}[P\{W>t_n(x_1-c_{2j},x_2-c_{1j},...,x_n-c_{1j})\}+P\{W>t_n(x_1-c_{1j},x_2-c_{2j},...,x_n-c_{1j})\}$$
$$+...+P\{W>t_n(x_1-c_{1j},...,x_n-c_{2j})\}-nP\{W>t_n(x_1-c_{1j},...,x_n-c_{1j})\}] \quad (3)$$

where W is a random variable with a Student's t distribution with $n-2k-1$ degrees of freedom. Hence, with this approximation, we transfer computations under the Gaussian Mixture Model $\mathbf{F}$ to computations of a Student's t distribution.

4 Simulations

If we consider a $N((1-\epsilon)\theta_0+\epsilon\theta,1)$ model as distribution G in the von Mises approximation (3), we observe in Fig. 1 that this approximation (dotted) is accurate considering $n=10, \theta_0=0, \epsilon=0.05, \alpha=0.1, \theta=1$, only with $k=20$ iterations and a simulation of $B=70$ replications in the computations of the simulated distribution of T_n.

In Fig. 2 we see that, even in the case that the underlying model is a $(1-\epsilon)N(\theta_0,1)+\epsilon N(\theta,1)$, the approximation is also accurate with the same values in the parameters as before.

References

1. Bickel, P.J.: On Some Robust Estimates of Location. Ann. Math. Statist. 36, 847–858 (1965)
2. Easton, G.S., Ronchetti, E.: General Saddlepoint Approximations with Applications to L Statistics. J. Amer. Statist. Assoc. 81, 420–430 (1986)
3. García-Pérez, A.: Von Mises Approximation of the Critical Value of a Test. Test 12, 385–411 (2003)
4. García-Pérez, A.: Chi-Square Tests under Models Close to the Normal Distribution. Metrika 63, 343–354 (2006)
5. García-Pérez, A.: *t*-tests with Models Close to the Normal Distribution. In: Balakrishnan, N., Castillo, E., Sarabia, J.M. (eds.) Advances in Distribution Theory, Order Statistics, and Inference, pp. 363–379. Birkhäuser-Springer, Boston (2006)
6. García-Pérez, A.: Approximations for F-tests which are Ratios of Sums of Squares of Independent Variables with a Model Close to the Normal. Test 17, 350–369 (2008)
7. García-Pérez, A.: Hotelling's T^2-Test with Multivariate Normal Mixture Populations: Approximations and Robustness. In: Pardo, L., Balakrishnan, N., Gil, M.Á. (eds.) Modern Mathematical Tools and Techniques in Capturing Complexity. Understanding Complex Systems, vol. 9, pp. 437–452. Springer, Heidelberg (2011)
8. García-Pérez, A.: Another Look at the Tail Area Influence Function. Metrika 73, 77–92 (2011)
9. García-Pérez, A.: A Linear Approximation to the Power Function of a Test. Metrika 75, 855–875 (2012)
10. Hall, P., Padmanabhan, A.P.: On the Bootstrap and the Trimmed Mean. J. Multivariate Anal. 41, 132–153 (1992)
11. Helmers, R., Jing, B.-Y., Qin, G., Zhou, W.: Saddlepoint Approximations to the Trimmed Mean. Bernoulli 10, 465–501 (2004)
12. Lehmann, E.L.: Theory of Point Estimation. John Wiley and Sons (1983)
13. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. John Wiley and Sons (2006)
14. Staudte, R.G., Sheather, A.J.: Robust Estimation and Testing. John Wiley and Sons (1990)
15. Stigler, S.M.: The Asymptotic Distribution of the Trimmed Mean. Ann. Stat. 1, 472–477 (1973)
16. Tukey, J.W., McLaughlin, D.H.: Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: trimming/winsorization 1. Sankhya A 25, 331–352 (1963)
17. Withers, C.S.: Expansions for the Distribution and Quantiles of a Regular Functional of the Empirical Distribution with Applications to Nonparametric Confidence Intervals. Ann. Stat. 11, 577–587 (1983)

Empirical Sensitivity Analysis on the Influence of the Shape of Fuzzy Data on the Estimation of Some Statistical Measures

María Asunción Lubiano[1], Sara de la Rosa de Sáa[1],
Beatriz Sinova[1,2], and María Ángeles Gil[1]

[1] Departamento de Estadística e I.O. y D.M.,
Universidad de Oviedo, 33007 Oviedo, Spain
{lubiano,delarosasara,sinovabeatriz,magil}@uniovi.es

[2] Department of Applied Mathematics, Computer Science and Statistics,
Ghent University, 9000 Gent, Belgium

Abstract. This paper means an introduction to analyze whether the choice of the shape for fuzzy data in their statistical analysis can or cannot affect the conclusions of such an analysis. More concretely, samples of fuzzy data are simulated in accordance with different assumptions (distributions) concerning four relevant points (namely, those determining their core and support), and later, by preserving core and support, the 'arms' are changed by considering trapezoidal, *Π*-curves, and some *LR* fuzzy numbers. For the simulations obtained with each of the considered shapes, several characteristics have been estimated: Aumann-type mean, 1-norm and wabl/ldev/rdev medians and Fréchet's variance. A comparative analysis with the bias, mean squared distance and variance of the estimates is finally included.

Keywords: fuzzy data, estimation, statistical measures, sensitivity analysis.

1 Introduction

Along the last years a distance-based methodology has been developed to analyze fuzzy number-valued data from a statistical perspective (see Blanco-Fernández *et al.* [2] for a recent review). The methodology assumes that data are generated from random elements taking on fuzzy numbers values (the so-called random fuzzy numbers or -one dimensional- fuzzy random variables in Puri and Ralescu's sense [10]).

Almost all the already developed methods refer to the estimation or to the hypothesis testing about some summary measures of the distributions of the random elements producing fuzzy-valued data. These methods are mostly theoretically supported, but empirical studies have been also conducted either to corroborate some of their generally stated properties or as an alternative when formal general results or conclusions cannot be stated.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_15

Most of these empirical developments have been based on simulations from random mechanisms leading to trapezoidal fuzzy number values. This assumption is often considered in practice to ease both the drawing and the computing processes (see, as a recent example the studies in De la Rosa de Sáa *et al.* [4]) although this is not at all mandatory from a formal viewpoint. Actually, Pedrycz [9], Grzegorzewski [6], [7], Grzegorzewski and Pasternak-Winiarska [8], Ban *et al.* [1], and others, have provided with different arguments to employ triangular or trapezoidal fuzzy numbers or approximations preserving ambiguity, expected interval, and so on.

An open problem that has been often commented in the papers related to the aforementioned distance-based methodology is that of discussing whether or not the shape of the fuzzy data influences the statistical conclusions. Since fuzzy data are essentially subjective in this respect, it is convenient to know whether this subjectivity can importantly affect the outputs from the methods.

This paper aims to analyze such a possible influence in which concerns the estimation of some summary measures, namely, three location ones (Aumann-type mean, and two L^1-type medians), and the Fréchet variance of the fuzzy dataset. For this purpose, simulations have been carried out from random mechanisms generating different types of fuzzy values, but data of different type sharing the core (i.e., the 1-level) and the closure of the support (0-level).

2 The Simulation Procedures

To analyze how sensitive the considered summary measures are w.r.t. changes in shape, the simulations we have carried out refer to the four key points characterizing the involved fuzzy numbers (more concretely, those determining their core and support). Six different shapes (T1 to T6, see Figure 1) based on the same four-tuple are separately employed. It is known that for any fuzzy number A there exist four numbers $a_1, a_2, a_3, a_4 \in \mathbb{R}$ and two functions $l_A, r_A : \mathbb{R} \to [0,1]$, where l_A is nondecreasing and r_A is nonincreasing, such that we can describe A with its membership function in the following manner,

$$A(x) = \begin{cases} 0 & \text{if } x < a_1 \\ l_A(x) & \text{if } a_1 \leq x < a_2 \\ 1 & \text{if } a_2 \leq x \leq a_3 \\ r_A(x) & \text{if } a_3 < x \leq a_4 \\ 0 & \text{if } a_4 < x. \end{cases}$$

The corresponding fuzzy numbers have been obtained by using different l_A and r_A functions: linear functions in T1 (trapezoidal fuzzy numbers), quadratic functions with T2 (Π-curves, see, for instance, [3]) and shape functions handling parametric monotonic Hermite-type interpolation in T3-T4 (LR fuzzy numbers using (2,2)-rational splines) and T5-T6 (LR fuzzy numbers using mixed exponential splines). For more details about the considered LR fuzzy numbers see, for instance, [13].

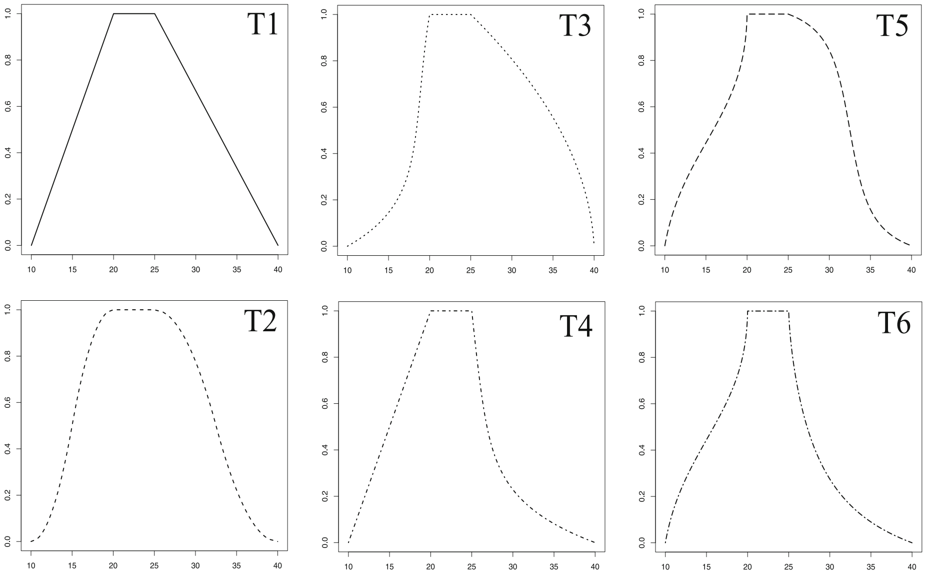


Fig. 1. Six types of fuzzy numbers sharing core and support and differing in shape. On the left, trapezoidal (top) and *Π*-curve (bottom), along with four different *LR* fuzzy numbers on the middle and the right

For each of these six shapes, some simulations studies have been conducted, generating the corresponding fuzzy numbers in two different ways:

Step 1. A sample of fuzzy numbers of the given shape has been obtained by simulating from

- four real-valued random variables X_i $(i = 1, 2, 3, 4)$, defining a random fuzzy number $\mathcal{X}$ in Puri and Ralescu's sense, namely, $X_1 = (\inf \mathcal{X}_1 + \sup \mathcal{X}_1)/2$, $X_2 = (\sup \mathcal{X}_1 - \inf \mathcal{X}_1)/2$, $X_3 = \inf \mathcal{X}_1 - \inf \mathcal{X}_0$, $X_4 = \sup \mathcal{X}_0 - \sup \mathcal{X}_1$ (whence $\inf \mathcal{X}_0 = X_1 - X_2 - X_3$, $\inf \mathcal{X}_1 = X_1 - X_2$, $\sup \mathcal{X}_1 = X_1 + X_2$, $\sup \mathcal{X}_0 = X_1 + X_2 + X_4$);
- In the FIRST STUDY (similar to some ones considered by Sinova *et al.*, see [11], [12]), the sample size is $n = 100$ and two cases related to these four random variables X_i have been considered: one in which X_i are independent (CASE 1) and another one in which they are dependent (CASE 2). More specifically, CASE 1 assumes that
  - •• $X_1 \sim \mathcal{N}(0, 1)$ and $X_2, X_3, X_4 \sim \chi^2_1$, all of them being independent

  whereas CASE 2 assumes that
  - •• $X_1 \sim \mathcal{N}(0, 1)$ and $X_2, X_3, X_4 \sim 1/(X_1^2 + 1)^2 + 0.1 \cdot \chi^2_1$, where χ^2_1 is supposed to be independent of X_1, and the three involved χ^2_1 being independent.
- In the SECOND STUDY (which follows the ideas by De la Rosa de Sáa *et al.* [4] in developing comparative studies in connection with questionnaires based on the fuzzy rating scale, using the referential [0,10]), the

simulation strategy has mimicked the human behavior by considering a finite mixture of three different procedures. Concretely, 100000 fuzzy values have been generated in the following way:

- ◦ 5% of the data have been obtained by first considering a simulation from a simple random sample of size 4 (X_1, X_2, X_3, X_4) from a beta population $X \sim \beta(1,1)$, later scaling it in [0,10] and finally considering the ordered sample $(X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)})$.
- ◦ 35% of the data have been obtained considering a simulation of four random variables X_i as follows:
  $$X_1 \sim \beta(1,1),$$
  $$X_2 \sim \text{Uniform}\big[0, \min\{1/10, X_1, 1-X_1\}\big],$$
  $$X_3 \sim \text{Uniform}\big[0, \min\{1/5, X_1 - X_2\}\big],$$
  $$X_4 \sim \text{Uniform}\big[0, \min\{1/5, 1 - X_1 - X_2\}\big];$$
- ◦ 60% of the data have been obtained considering a simulation of four random variables X_i as follows:
  $$X_1 \sim \beta(1,1),$$
  $$X_2 \sim \begin{cases} \text{Exp}(200) & \text{if } X_1 \in [0.25, 0.75] \\ \text{Exp}(100 + 4\,X_1) & \text{if } X_1 < 0.25 \\ \text{Exp}(500 - 4\,X_1) & \text{otherwise} \end{cases}$$
  $$X_3 \sim \begin{cases} \gamma(4,100) & \text{if } X_1 - X_2 \geq 0.25 \\ \gamma(4, 100 + 4\,X_1) & \text{otherwise} \end{cases}$$
  $$X_4 \sim \begin{cases} \gamma(4,100) & \text{if } 1 - X_1 - X_2 \leq 0.75 \\ \gamma(4, 500 - 4\,X_1) & \text{otherwise.} \end{cases}$$

Step 2. $N = 1000$ replications of *Step 1* in the first study have been considered and the 100000 fuzzy values from the second study have been divided randomly (and without replacement) into 1000 samples of size $n = 100$. So in both studies, there are 1000 available samples of size $n = 100$.

Step 3. The population summary measures have been approximated on the basis of 35.000 replications.

Step 4. The estimates have been complemented with the average distance-based bias along the 1000 samples, and some other associated mean errors.

Distances have been computed by considering three different metrics: the L^2 metric ρ_2, the L^1 metric ρ_1 (see Diamond and Kloeden [5]) and the L^1 metric $\mathscr{D}_1$ (a particular case of that introduced by Sinova *et al.* [11]), where for fuzzy numbers $\widetilde{U}, \widetilde{V}$ they are given by

$$\rho_2(\widetilde{U},\widetilde{V}) = \sqrt{\frac{1}{2}\int_{[0,1]} \Big[(\inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha)^2 + (\sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha)^2\Big]\, d\alpha},$$

$$\rho_1(\widetilde{U},\widetilde{V}) = \frac{1}{2}\int_{[0,1]} \Big[|\inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha| + |\sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha|\Big]\, d\alpha,$$

$$\mathscr{D}_1(\widetilde{U},\widetilde{V}) = |\text{wabl}(\widetilde{U}) - \text{wabl}(\widetilde{U})|$$
$$+\frac{1}{2}\int_{[0,1]} \Big[|\text{ldev}\, \widetilde{U}_\alpha - \text{ldev}\, \widetilde{V}_\alpha| + |\text{rdev}\, \widetilde{U}_\alpha - \text{rdev}\, \widetilde{V}_\alpha|\Big]\, d\alpha,$$

with $\text{wabl}(\widetilde{U}) = \int_{[0,1]} (\inf \widetilde{U}_\alpha + \sup \widetilde{U}_\alpha)\, d\alpha/2$, $\text{ldev}\, \widetilde{U}_\alpha = \text{wabl}(\widetilde{U}) - \inf \widetilde{U}_\alpha$, $\text{rdev}\, \widetilde{U}_\alpha = \sup \widetilde{U}_\alpha - \text{wabl}(\widetilde{U})$.

The outputs for this first simulation study have been collected in Table 1 for the mean errors in estimating the summary measures and in Figure 2 and Table 2 for their estimates.

Table 1. Mean errors in the estimation of some summary measures with the first simulations (CASES 1 and 2) for the six different types of fuzzy numbers in Figure 1

CASE 1

		ρ_2-Mean			ρ_1-Median			$\mathscr{D}_1$-Median			ρ_2-Variance		
Type	Error	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$
T1	Bias	0.004	0.004	0.005	0.008	0.008	0.009	0.007	0.008	0.011	0.028	0.028	0.028
	Variance	0.028	0.035	0.055	0.027	0.036	0.056	0.029	0.038	0.058	0.583	0.583	0.583
	MSE	0.028	0.035	0.055	0.028	0.036	0.056	0.029	0.038	0.058	0.584	0.584	0.584
T2	Bias	0.004	0.004	0.005	0.008	0.008	0.009	0.007	0.008	0.010	0.028	0.028	0.028
	Variance	0.028	0.034	0.054	0.027	0.035	0.055	0.029	0.037	0.058	0.576	0.576	0.576
	MSE	0.028	0.034	0.054	0.027	0.035	0.056	0.029	0.037	0.058	0.576	0.576	0.576
T3	Bias	0.005	0.005	0.005	0.008	0.008	0.008	0.008	0.008	0.011	0.022	0.022	0.022
	Variance	0.028	0.035	0.055	0.027	0.036	0.056	0.029	0.038	0.059	0.583	0.583	0.583
	MSE	0.028	0.035	0.055	0.027	0.036	0.056	0.030	0.038	0.059	0.583	0.583	0.583
T4	Bias	0.004	0.004	0.004	0.007	0.007	0.009	0.008	0.008	0.008	0.030	0.030	0.030
	Variance	0.027	0.033	0.051	0.026	0.034	0.053	0.026	0.034	0.051	0.562	0.562	0.562
	MSE	0.027	0.033	0.051	0.026	0.034	0.053	0.026	0.034	0.051	0.563	0.563	0.563
T5	Bias	0.004	0.004	0.005	0.008	0.008	0.009	0.008	0.008	0.010	0.027	0.027	0.027
	Variance	0.028	0.035	0.054	0.027	0.035	0.055	0.029	0.037	0.058	0.574	0.574	0.574
	MSE	0.028	0.035	0.054	0.027	0.035	0.055	0.029	0.037	0.058	0.575	0.575	0.575
T6	Bias	0.004	0.004	0.004	0.007	0.008	0.009	0.008	0.008	0.009	0.029	0.029	0.029
	Variance	0.027	0.033	0.051	0.026	0.034	0.052	0.026	0.033	0.051	0.558	0.558	0.558
	MSE	0.027	0.033	0.051	0.026	0.034	0.052	0.026	0.034	0.051	0.559	0.559	0.559

CASE 2

		ρ_2-Mean			ρ_1-Median			$\mathscr{D}_1$-Median			ρ_2-Variance		
Type	Error	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$
T1	Bias	0.003	0.003	0.003	0.002	0.002	0.002	0.005	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.020	0.005	0.007	0.012	0.021	0.026	0.041	0.023	0.023	0.023
	MSE	0.011	0.013	0.020	0.005	0.007	0.012	0.021	0.026	0.041	0.023	0.023	0.023
T2	Bias	0.003	0.003	0.003	0.002	0.002	0.002	0.005	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.020	0.005	0.006	0.011	0.021	0.026	0.041	0.023	0.023	0.023
	MSE	0.011	0.013	0.020	0.005	0.006	0.011	0.021	0.026	0.041	0.023	0.023	0.023
T3	Bias	0.003	0.003	0.003	0.001	0.002	0.002	0.005	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.019	0.006	0.008	0.014	0.021	0.025	0.040	0.023	0.023	0.023
	MSE	0.011	0.013	0.019	0.006	0.008	0.014	0.021	0.025	0.040	0.023	0.023	0.023
T4	Bias	0.003	0.003	0.003	0.002	0.002	0.003	0.004	0.004	0.005	0.004	0.004	0.004
	Variance	0.011	0.013	0.019	0.007	0.009	0.015	0.020	0.024	0.038	0.023	0.023	0.023
	MSE	0.011	0.013	0.019	0.007	0.009	0.015	0.020	0.024	0.038	0.023	0.023	0.023
T5	Bias	0.003	0.003	0.003	0.002	0.002	0.002	0.004	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.019	0.005	0.007	0.012	0.021	0.025	0.040	0.023	0.023	0.023
	MSE	0.011	0.013	0.019	0.005	0.007	0.012	0.021	0.025	0.040	0.023	0.023	0.023
T6	Bias	0.003	0.003	0.003	0.002	0.002	0.003	0.004	0.004	0.005	0.004	0.004	0.004
	Variance	0.011	0.012	0.018	0.007	0.009	0.016	0.020	0.024	0.037	0.022	0.022	0.022
	MSE	0.011	0.012	0.018	0.007	0.009	0.016	0.020	0.024	0.037	0.023	0.023	0.023

On the basis of the outputs in Table 1 one can empirically conclude to some extent that the shape of the considered data scarcely affects the bias, variance and mean squared error of the summary measures estimates.

Table 2. Monte Carlo estimate of the Fréchet ρ_2-variance in the first simulations

Variance	T1	T2	T3	T4	T5	T6
CASE 1	3.629	3.547	3.587	3.402	3.565	3.402
CASE 2	1.268	1.262	1.254	1.223	1.258	1.219

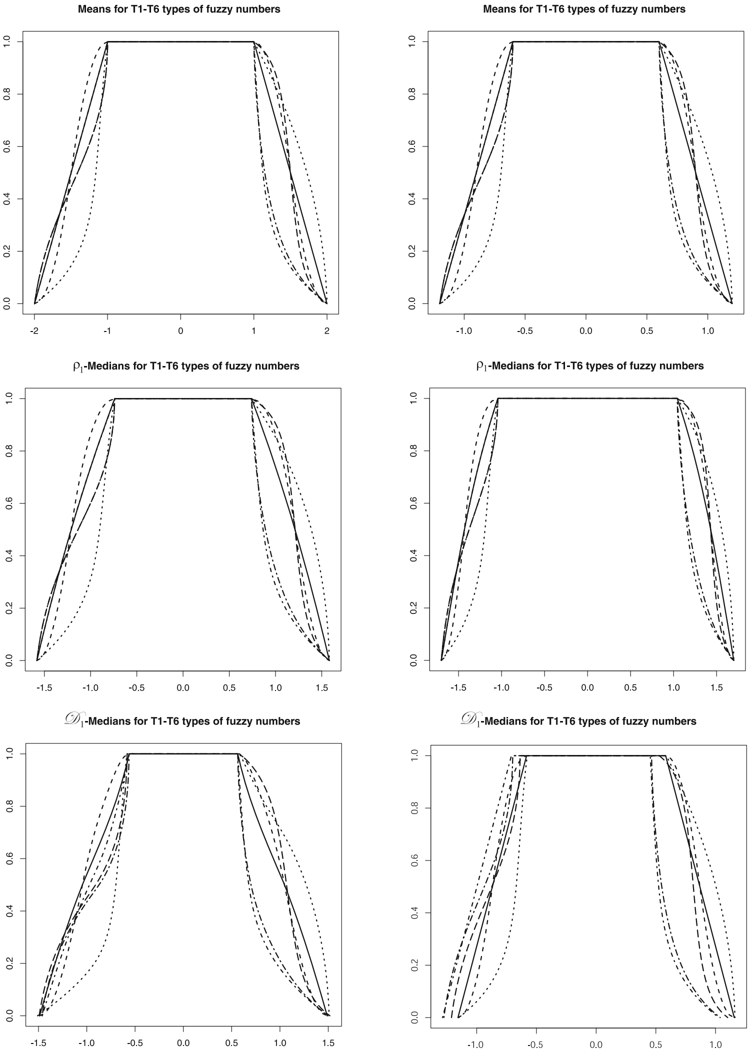


Fig. 2. Monte Carlo estimates of the (Aumann type) means and ρ_1- and $\mathscr{D}_1$-medians in CASE 1 (on the left) and CASE 2 (on the right) of the first simulations

The same happens for the estimates of the ρ_2-Fréchet variance in Table 2. The estimates of the location measures, graphically displayed in Figure 2, are more influenced by the shape of the involved fuzzy data. Nevertheless, the location estimates are indeed closer than the original data.

The outputs for the second simulation study have been collected in Table 3 for the mean errors in estimating the summary measures and in Figure 3 and Table 4 for their estimates. On the basis of the outputs in Table 3 one can empirically conclude to some extent that the shape of the considered data does not strongly affect the bias, variance and mean squared error of the summary measures estimates.

Table 3. Mean errors in the estimation of the summary measures with the second simulations for the six different types of fuzzy numbers in Figure 1

		ρ_2-Mean			ρ_1-Median			$\mathscr{D}_1$-Median			ρ_2-Variance		
Type	Error	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$
T1	Bias	0.002	0.002	0.003	0.009	0.009	0.012	0.005	0.005	0.005	0.016	0.016	0.016
	Variance	0.083	0.086	0.106	0.220	0.238	0.323	0.213	0.217	0.252	0.523	0.523	0.523
	MSE	0.083	0.086	0.106	0.220	0.238	0.323	0.213	0.217	0.252	0.524	0.524	0.524
T2	Bias	0.002	0.002	0.003	0.008	0.009	0.012	0.005	0.005	0.005	0.016	0.016	0.016
	Variance	0.083	0.085	0.105	0.220	0.237	0.322	0.213	0.217	0.253	0.524	0.524	0.524
	MSE	0.083	0.085	0.105	0.221	0.237	0.322	0.213	0.217	0.253	0.525	0.525	0.525
T3	Bias	0.002	0.002	0.002	0.010	0.011	0.013	0.005	0.005	0.006	0.018	0.018	0.018
	Variance	0.083	0.086	0.105	0.219	0.236	0.317	0.211	0.214	0.248	0.533	0.533	0.533
	MSE	0.083	0.086	0.105	0.219	0.236	0.318	0.211	0.214	0.248	0.534	0.534	0.534
T4	Bias	0.002	0.002	0.003	0.007	0.008	0.009	0.002	0.002	0.003	0.016	0.016	0.016
	Variance	0.084	0.086	0.105	0.218	0.234	0.316	0.219	0.221	0.254	0.537	0.537	0.537
	MSE	0.084	0.086	0.105	0.218	0.235	0.316	0.219	0.221	0.254	0.537	0.537	0.537
T5	Bias	0.002	0.002	0.003	0.009	0.009	0.012	0.004	0.004	0.004	0.017	0.017	0.017
	Variance	0.083	0.086	0.106	0.219	0.236	0.320	0.214	0.217	0.252	0.527	0.527	0.527
	MSE	0.083	0.086	0.106	0.219	0.237	0.320	0.214	0.217	0.252	0.527	0.527	0.527
T6	Bias	0.002	0.002	0.003	0.008	0.008	0.009	0.001	0.001	0.002	0.016	0.016	0.016
	Variance	0.084	0.086	0.105	0.217	0.234	0.315	0.218	0.221	0.253	0.539	0.539	0.539
	MSE	0.084	0.086	0.105	0.217	0.234	0.315	0.218	0.221	0.253	0.539	0.539	0.539

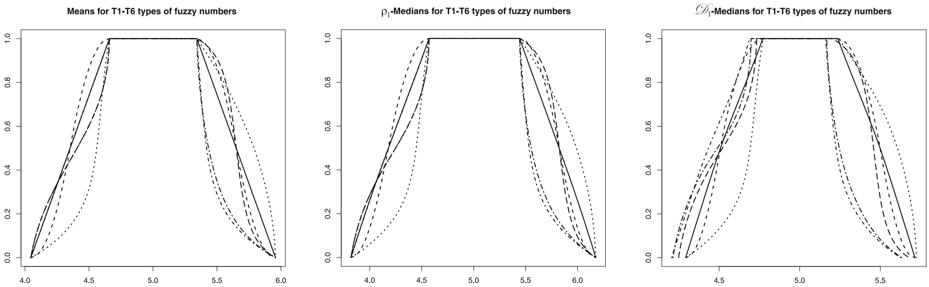


Fig. 3. Approximated estimates of the (Aumann type) means and ρ_1- and $\mathscr{D}_1$-medians for the second simulations

Table 4. Approximated estimate of the Fréchet ρ_2-variance in the second simulations

	T1	T2	T3	T4	T5	T6
Variance	7.921	7.902	7.950	7.971	7.926	7.983

The same happens for the estimates of the ρ_2-Fréchet variance in Table 4, although the shape difference influences slightly more than for the first study. The estimates of the location measures, graphically displayed in Figure 3, are more influenced by the shape of the involved fuzzy data, also slightly more than for the first simulations. Again, the location estimates are indeed closer than the original data.

As a clear extension of the study in this paper, it is a must to develop comparison concerning the influence on the power of hypothesis testing involving fuzzy data.

Acknowledgments. The research in this paper has been partially supported by the Severo Ochoa Grant BP12012 from the Principality of Asturias (de la Rosa de Sáa) and the FPU Grant AP2009-1197 from the Spanish Ministry of Education (Sinova). The support is gratefully acknowledged.

References

1. Ban, A., Coroianu, L., Grzegorzewski, P.: Trapezoidal approximation and aggregation. Fuzzy Sets Syst. 177, 45–59 (2011)
2. Blanco-Fernández, A., Casals, M.R., Colubi, A., Corral, N., García-Bárzana, M., Gil, M.A., González-Rodríguez, G., López, M.T., Lubiano, M.A., Montenegro, M., Ramos-Guajardo, A.B., De la Rosa de Sáa, S., Sinova, B.: A distance-based statistical analysis of fuzzy number-valued data. Int. J. Approx. Reas (2014), doi:10.1016/j.ijar.2013.09.020
3. Cox, E.: The fuzzy Systems Handbook. Academic Press, Cambridge (1994)
4. De la Rosa de Sáa, S., Gil, M.A., González-Rodríguez, G., López, M.T., Lubiano, M.A.: Fuzzy rating scale-based questionnaires and their statistical analysis. IEEE Trans. Fuzzy Syst (2014), doi:10.1109/TFUZZ.2014.2307895
5. Diamond, P., Kloeden, P.: Metric spaces of fuzzy sets. Fuzzy Sets Syst. 35, 241–249 (1990)
6. Grzegorzewski, P.: Trapezoidal approximations of fuzzy numbers preserving the expected interval - algorithms and properties. Fuzzy Sets Syst. 159, 1354–1364 (2008)
7. Grzegorzewski, P.: Fuzzy number approximation via shadowed sets. Inform. Sci. 225, 35–46 (2013)
8. Grzegorzewski, P., Pasternak-Winiarska, K.: Trapezoidal approximations of fuzzy numbers with restrictions on the support and core. In: Proc. 7th Conf. EUSFLAT-2011 and LFA-2011, pp. 749–756. Atlantis Press, Paris (2011)
9. Pedrycz, W.: Why triangular membership functions? Fuzzy Sets Syst. 64(1), 21–30 (1994)

10. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
11. Sinova, B., De la Rosa de Sáa, S., Gil, M.A.: A generalized L^1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data. Inform. Sci. 242, 22–34 (2013)
12. Sinova, B., Gil, M.A., Colubi, A., Van Aelst, S.: The median of a random fuzzy number. The 1-norm distance approach. Fuzzy Sets Syst. 200, 99–115 (2012)
13. Stefanini, L., Sorini, L., Guerra, M.L.: Parametric representation of fuzzy numbers and applications to fuzzy calulus. Fuzzy Sets Syst. 157, 2423–2455 (2006)

On the Robustness of Absolute Deviations with Fuzzy Data

Sara de la Rosa de Sáa[1], Peter Filzmoser[2],
María Ángeles Gil[1], and María Asunción Lubiano[1]

[1] Departamento de Estadística e I.O. y D.M.,
Universidad de Oviedo, 33007 Oviedo, Spain
{delarosasara,magil,lubiano}@uniovi.es

[2] Institut für Statistik und Wahrscheinlichkeitstheorie,
Technische Universität Wien, 1040 Wien, Austria
P.Filzmoser@tuwien.ac.at

Abstract. Often atypical observations separated from the majority or deviate from the general pattern appear in the datasets. Classical estimators such as the sample mean or the sample variance, can be substantially affected by these observations, which are referred to as outliers. Robust statistics provides methods which are not unduly influenced by atypical data.

In this paper an introductory empirical study is developed to compare the robustness of the scale estimator 'Median Absolute Deviation' in contrast to the classical scale estimator 'Average Absolute Deviation' in a fuzzy setting. Both estimators are defined on the basis of the Aumann-type mean, the 1-norm median for random fuzzy numbers along with an L^1-type metric between fuzzy numbers, and some of their properties are examined. Outliers will be introduced in simulated fuzzy data to analyze how much these two estimators are influenced by them.

Keywords: fuzzy data, robustness, median absolute deviation, average absolute deviation.

1 Introduction

Over the last years, the consideration of different sources of imprecision for generating and modelling experimental data have implied the development of advanced statistical and soft computing techniques capable to cope with this kind of data.

By combining probabilistic uncertainty with fuzzy imprecision, the concept of the so-called random fuzzy numbers (RFNs for short, see [10]) arises as a suitable and well-formalized model within the probabilistic setting. In the study of RFNs we could describe their behaviour by means of certain measures summarizing the central tendency or location. Some of the most relevant measures in this context are the mean and the median (see, for instance, [1,3,4,6,9,13]).

Another useful summary tool associated with the distribution of random magnitudes is the measurement of the dispersion, given that when there is no variation of observations the statistical methodology is not of interest. Measures of

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_16

variability can be used in practice as descriptive statistics for data analysis, to compare the dispersion of different datasets, to formulate rules for the detection of outliers, and so on. In the fuzzy setting the variability of the values of an RFN have been measured by means of the Fréchet variance defined in terms of an L^2-type metric (see, for instance, [7,11]).

In statistical applications, it is very valuable for estimators of summary measures to be robust. In this paper, we refer with robust statistics to statistical approaches that are less influenced either by outlying observations or by deviations from strict statistical model assumptions (Maronna *et al.* [8]). A different method for quantifying the robustness that does not address here has been proposed by Zieliński [16].

The Aumann-type sample mean is highly affected by the presence of outliers. Sinova *et al.* [13] have shown that the sample 1-norm median is a good estimate of the location for fuzzy numbers and it is a robust alternative to the sample mean in estimating location.

Similarly, the Fréchet variance can be very adversely influenced by atypical values. The aim of this contribution is to introduce scale measures for RFNs on the basis of an L^1-type metric and to analyze their sensitivity to either changes of values or presence of outliers.

In Section 2 the preliminaries on fuzzy numbers, arithmetic and metrics between them, and the concept of random fuzzy number and their summary measures of central tendency will be established. Section 3 introduces some scale measures and analyzes some relevant properties. In Section 4 the influence of outliers on these measures is illustrated by means of a simulation study. Finally, some future research directions will be commented.

2 Preliminaries about Fuzzy Statistics

Let $\mathcal{F}_c^*(\mathbb{R})$ be the space of bounded fuzzy numbers. A (bounded) *fuzzy number* $\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R})$ is an ill-defined quantity or value which can be formally characterized by means of a mapping $\widetilde{U} : \mathbb{R} \to [0,1]$ such that the *α-level* set of $\widetilde{U}$, defined as $\widetilde{U}_\alpha = \{\mathbf{x} \in \mathbb{R} : \widetilde{U}(\mathbf{x}) \geq \alpha\}$ if $\alpha \in (0,1]$, and as the closure of the support of $\widetilde{U}$ if $\alpha = 0$, is a nonempty compact interval. For each $x \in \mathbb{R}$, the value $\widetilde{U}(x)$ can be interpreted as the 'degree of compatibility' of x with the property 'defining' $\widetilde{U}$.

On the space $\mathcal{F}_c^*(\mathbb{R})$ one can consider the usual ***fuzzy arithmetic*** based on Zadeh's extension principle [15], which coincides level-wise with the usual interval arithmetic. Concretely the operations required for the statistical analysis of fuzzy data are the sum of fuzzy numbers and the product of fuzzy numbers by a scalar satisfying that

$$(\widetilde{U}+\widetilde{V})_\alpha = \widetilde{U}_\alpha+\widetilde{V}_\alpha = \{u+v \mid u \in \widetilde{U}_\alpha,\, v \in \widetilde{V}_\alpha\}, \quad (\gamma\widetilde{U})_\alpha = \gamma\widetilde{U}_\alpha = \{\gamma u \mid u \in \widetilde{U}_\alpha\}$$

for all $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c^*(\mathbb{R})$, $\gamma \in \mathbb{R}$ and $\alpha \in [0,1]$.

$(\mathcal{F}_c^*(\mathbb{R}), +, \cdot)$ is not a linear but a semilinear space (since there is no inverse element for the sum). This lack of a suitable definition for the difference between two fuzzy numbers is often overcome by considering *metrics between fuzzy numbers* which are intuitive, versatile and easy-to-use.

In this article we consider the L^1 metric by Diamond and Kloeden [2], which extends the Euclidean metric in $\mathbb{R}$ and Vitale's L^1 metric between nonempty compact intervals [14]. Given $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c^*(\mathbb{R})$, the mapping $\rho_1 : \mathcal{F}_c^*(\mathbb{R}) \times \mathcal{F}_c^*(\mathbb{R}) \to [0, +\infty)$ defined as

$$\rho_1(\widetilde{U}, \widetilde{V}) = \frac{1}{2} \int_{(0,1]} \left(\left| \inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha \right| + \left| \sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha \right| \right) d\alpha$$

is to be referred as the *1-norm distance* between two fuzzy numbers.

To formalize the random mechanisms that produce fuzzy data, the *random fuzzy numbers* (Puri and Ralescu [10]) will be considered.

Definition 1. *Given a probability space $(\Omega, \mathcal{A}, P)$, an associated* ***random fuzzy number*** *(RFN) is a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R})$ such that for all $\alpha \in [0,1]$ the α-level mapping $\mathcal{X}_\alpha : \Omega \to \mathcal{P}(\mathbb{R})$ (with $\mathcal{X}_\alpha(\omega) = \big(\mathcal{X}(\omega)\big)_\alpha$) is a compact random interval (that is, for all $\alpha \in [0,1]$ the real-valued mappings* $\inf \mathcal{X}_\alpha$ *and* $\sup \mathcal{X}_\alpha$ *are random variables).*

This notion can be equivalently formalized as a Borel-measurable mapping w.r.t. the Borel σ-field generated on $\mathcal{F}_c^*(\mathbb{R})$ by the topology induced by the metric ρ_1. Consequently, one can properly refer to the induced distribution of an RFN, the independence of two RFNs, etc.

Definition 2. *Let $\mathcal{X}$ be an RFN. The* ***Aumann-type mean*** *of $\mathcal{X}$ (defined by Puri and Ralescu [10]) is the fuzzy number $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c^*(\mathbb{R})$, if it exists, such that for each $\alpha \in [0,1]$*

$$\big(\widetilde{E}(\mathcal{X})\big)_\alpha = [E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)] .$$

Definition 3. *Let $\mathcal{X}$ be an RFN. The* ***1-norm median*** *of $\mathcal{X}$ (defined by Sinova* et al. *[13]) is the fuzzy number $\widetilde{\mathrm{Me}}(\mathcal{X}) \in \mathcal{F}_c^*(\mathbb{R})$, if it exists, such that for each $\alpha \in [0,1]$*

$$\left(\widetilde{\mathrm{Me}}(\mathcal{X})\right)_\alpha = \left[\mathrm{Me}\big(\inf \mathcal{X}_\alpha\big), \mathrm{Me}\big(\sup \mathcal{X}_\alpha\big)\right] ,$$

where in case $\mathrm{Me}\big(\inf \mathcal{X}_\alpha\big)$ *or* $\mathrm{Me}\big(\sup \mathcal{X}_\alpha\big)$ *are nonunique we will follow the most usual convention, that is, we will consider the midpoint of the interval of medians.*

Both the mean and the 1-norm median of a RFN preserve most of the basic properties of the mean and the median of random variables (e.g., they are equivariant by positive affine transformations).

3 Measures of Scale

Measures like the sample mean or the sample 1-norm median provide a good estimate of the central tendency or location of a dataset. Nevertheless, it is also very valuable to measure/estimate the variability of the data. Responsible for carrying out this task are the measures of scale.

In this section, two scale measures based on location measures (one of them on the mean and the other one on the 1-norm median), and on an L^1-type metric, are introduced for RFNs, and some first properties are stated.

Definition 4. *Let $\mathcal{X}$ be an RFN. The* ***median absolute deviation about the median*** *is the real number $\widetilde{\mathrm{MAD}}(\mathcal{X}) \in [0, \infty)$ defined as*

$$\widetilde{\mathrm{MAD}}(\mathcal{X}) = \mathrm{Me}\left(\rho_1(\mathcal{X}, \widetilde{\mathrm{Me}}(\mathcal{X}))\right).$$

Definition 5. *Let $\mathcal{X}$ be an RFN. The* ***average absolute deviation about the mean*** *is the real number $\widetilde{\mathrm{AAD}}(\mathcal{X}) \in [0, \infty)$ defined as*

$$\widetilde{\mathrm{AAD}}(\mathcal{X}) = E\left(\rho_1(\mathcal{X}, \widetilde{E}(\mathcal{X}))\right).$$

These measures satisfy the following Propositions:

Proposition 1. *$\widetilde{\mathrm{MAD}}$ satisfies the* ***shift invariance*** *and the* ***scale equivariance*** *conditions. That is, given $\gamma \in \mathbb{R}$, $\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R})$ and $\mathcal{X}$ an RFN, then:*

$$\widetilde{\mathrm{MAD}}(\gamma \cdot \mathcal{X} + \widetilde{U}) = |\gamma| \cdot \widetilde{\mathrm{MAD}}(\mathcal{X}).$$

Proof. Taking into account the equivariance under 'linear' transformations of the 1-norm median of an RFN, and considering that ρ_1 is a traslational invariant and equivariant under positive homotheties metric, along with the properties of the median for real-valued random variables, we have that:

$$\widetilde{\mathrm{MAD}}(\gamma \cdot \mathcal{X} + \widetilde{U}) = \mathrm{Me}\left(\rho_1(\gamma \cdot \mathcal{X} + \widetilde{U}, \widetilde{\mathrm{Me}}(\gamma \cdot \mathcal{X} + \widetilde{U}))\right)$$

$$= \mathrm{Me}\left(\rho_1(\gamma \cdot \mathcal{X} + \widetilde{U}, \gamma \cdot \widetilde{\mathrm{Me}}(\mathcal{X}) + \widetilde{U})\right) = \mathrm{Me}\left(|\gamma| \cdot \rho_1(\mathcal{X}, \widetilde{\mathrm{Me}}(\mathcal{X}))\right)$$

$$= |\gamma| \cdot \mathrm{Me}\left(\rho_1(\mathcal{X}, \widetilde{\mathrm{Me}}(\mathcal{X}))\right) = |\gamma| \cdot \widetilde{\mathrm{MAD}}(\mathcal{X}).$$

Proposition 2. *$\widetilde{\mathrm{AAD}}$ also satisfies the* ***shift invariance*** *and the* ***scale equivariance*** *conditions. Therefore, given $\gamma \in \mathbb{R}$, $\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R})$ and $\mathcal{X}$ an RFN, then:*

$$\widetilde{\mathrm{AAD}}(\gamma \cdot \mathcal{X} + \widetilde{U}) = |\gamma| \cdot \widetilde{\mathrm{AAD}}(\mathcal{X}).$$

4 Simulation Study

The simulation study presented in this section aims to empirically check the robustness of the estimator $\widehat{\widetilde{\mathrm{MAD}}}$ in comparison to the estimator $\widehat{\widetilde{\mathrm{AAD}}}$, which is not robust.

A popular and successful measure of the robustness of an estimator is its breakdown point, introduced by Hampel [5]. It is the minimum proportion of sample data which should be perturbed to get an arbitrary large or small estimator value. Thus, estimators with a high breakdown point are not influenced by a high amount of atypical data. Some of them can reach a breakdown point of 50%, the highest bound.

Following the definition of breakdown point for scale estimators by Rousseeuw and Croux [12], for any sample $\widetilde{\mathbf{x}}_n$ from an RFN $\mathcal{X}$, the so-called *finite sample breakdown point* (fsbp for short) of a scale estimator $\widehat{\widetilde{\mathrm{S}}(\mathcal{X})_n}$ is defined by:

$$\mathrm{fsbp}^*(\widehat{\widetilde{\mathrm{S}}(\mathcal{X})_n}, \widetilde{\mathbf{x}}_n) = \min\left\{\mathrm{fsbp}^+(\widehat{\widetilde{\mathrm{S}}(\mathcal{X})_n}, \widetilde{\mathbf{x}}_n),\ \mathrm{fsbp}^-(\widehat{\widetilde{\mathrm{S}}(\mathcal{X})_n}, \widetilde{\mathbf{x}}_n)\right\}$$

where

$$\mathrm{fsbp}^+(\widehat{\widetilde{\mathrm{S}}(\mathcal{X})_n}, \widetilde{\mathbf{x}}_n) = \min\left\{\frac{k}{n}; \sup_{\widetilde{\mathbf{y}}_{n,k}} \widehat{\widetilde{\mathrm{S}}(\widetilde{\mathbf{y}}_{n,k})} = \infty\right\}$$

and

$$\mathrm{fsbp}^-(\widehat{\widetilde{\mathrm{S}}(\mathcal{X})_n}, \widetilde{\mathbf{x}}_n) = \min\left\{\frac{k}{n}; \inf_{\widetilde{\mathbf{y}}_{n,k}} \widehat{\widetilde{\mathrm{S}}(\widetilde{\mathbf{y}}_{n,k})} = 0\right\}$$

with $\widetilde{\mathbf{y}}_{n,k}$ obtained by replacing any k observations of $\widetilde{\mathbf{x}}_n$ by arbitrary values. The quantities fsbp^+ and fsbp^- are called the *explosion breakdown point* and the *implosion breakdown point.*

Following Sinova *et al.* [13], the simulations have been performed by considering trapezoidal RFNs $\mathcal{X} = \mathrm{Tra}(\inf \mathcal{X}_0, \inf \mathcal{X}_1, \sup \mathcal{X}_1, \sup \mathcal{X}_0)$, each of them characterized by means of the following four real-valued random variables:

- $X_1 = (\inf \mathcal{X}_1 + \sup \mathcal{X}_1)/2$, $X_2 = (\sup \mathcal{X}_1 - \inf \mathcal{X}_1)/2$,
- $X_3 = \inf \mathcal{X}_1 - \inf \mathcal{X}_0$, $X_4 = \sup \mathcal{X}_0 - \sup \mathcal{X}_1$,

whence $\mathcal{X} = \mathrm{Tra}(X_1 - X_2 - X_3, X_1 - X_2, X_1 + X_2, X_1 + X_2 + X_4)$.

We have considered samples of size $n = 100$ which can be split into *non-contaminated subsamples* of size $n(1 - c_p)$ (where c_p denotes the proportion of contamination) and *contaminated subsamples* of size $n \cdot c_p$.

For the non-contaminated subsamples we have assumed that $X_1 \rightsquigarrow \mathcal{N}(0, 1)$, $X_2, X_3, X_4 \rightsquigarrow \chi^2_1$.

In the case of the contaminated subsamples two different cases have been studied:

- Explosion: $X_1 \rightsquigarrow \mathcal{N}(0,3)+100$ and $X_2, X_3, X_4 \rightsquigarrow \chi^2_4 + 100$
- Implosion: one datum of the non-contaminated sample has been randomly chosen and repeated $n \cdot c_p$ times.

For each sample we have computed the estimators $\widehat{\widetilde{\text{MAD}}}$ and $\widehat{\widetilde{\text{AAD}}}$ and we have considered 10000 replications. The results of the simulation are shown in Table 1 for the explosion and in Table 2 for the implosion. Each table entry is the mean of the estimators on the 10000 replications.

Table 1. Explosion: behaviour of the $\widehat{\widetilde{\text{MAD}}}$ and $\widehat{\widetilde{\text{AAD}}}$ when outliers are introduced in the sample

c_p	0	0.1	0.2	0.3	0.4	0.5
$\widehat{\widetilde{\text{MAD}}}$	1.06	1.2	0.42	1.79	2.58	76.88
$\widehat{\widetilde{\text{AAD}}}$	1.4	27.89	49.47	64.91	74.17	77.27

Table 2. Implosion: behaviour of the $\widehat{\widetilde{\text{MAD}}}$ and $\widehat{\widetilde{\text{AAD}}}$ when there are repeated data in the sample

c_p	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\widehat{\widetilde{\text{MAD}}}$	1.21	1.2	1.17	1.05	0.77	0.07	0	0	0	0	0
$\widehat{\widetilde{\text{AAD}}}$	1.4	1.4	1.37	1.33	1.25	1.14	1	0.82	0.59	0.32	0

In Figure 1 the explosion and implosion maxbias curves for the estimators $\widehat{\widetilde{\text{MAD}}}$ and $\widehat{\widetilde{\text{AAD}}}$ show the behaviour of these measures when perturbations are introduced progressively in the sample.

We can notice that the fsbp for the estimator $\widehat{\widetilde{\text{MAD}}}$ is 50%, since this is the smallest proportion of contamination that is needed to let $\widehat{\widetilde{\text{MAD}}}$ explode to infinity or implode to zero.

Conversely, we can deduce that the fsbp for the estimator $\widehat{\widetilde{\text{AAD}}}$ is 0%.

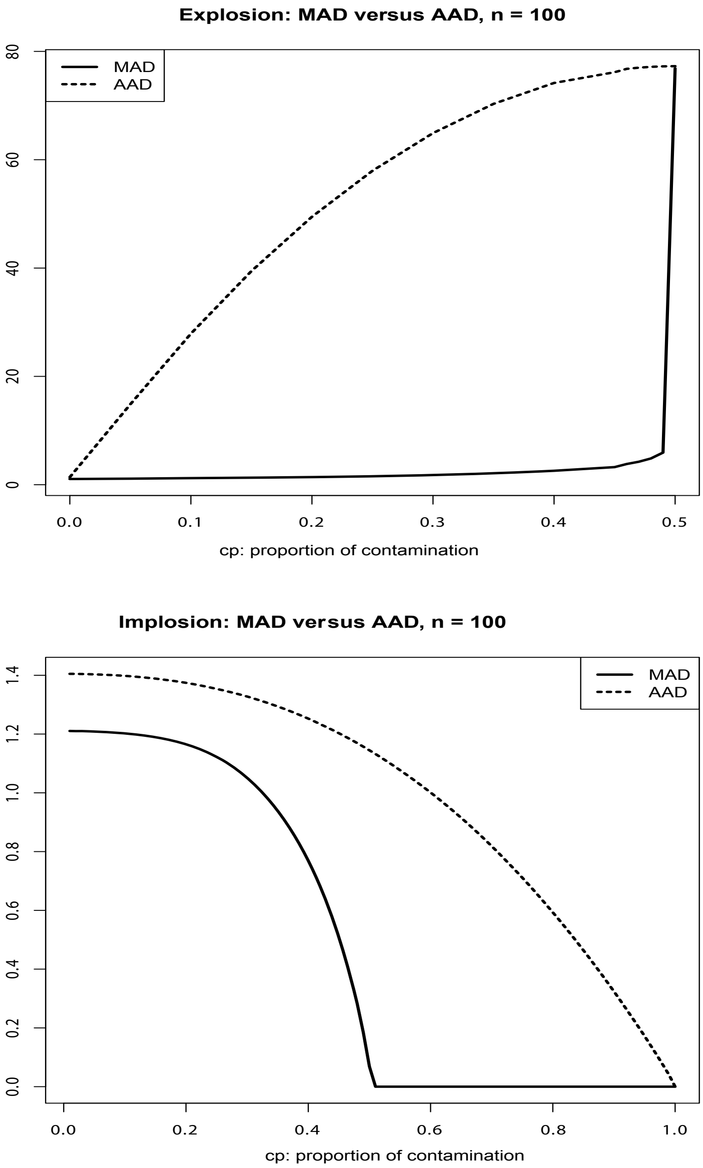


Fig. 1. Comparison of the effect of increasing contamination in the scale estimators $\widehat{\widetilde{\mathrm{MAD}}}$ and $\widehat{\mathrm{AAD}}$ through explosion and implosion maxbias curves

5 Concluding Remarks

In this paper the definition of the robust scale measure $\widetilde{\mathrm{MAD}}$ for fuzzy data has been introduced. Its property of equivariance under positive affine transformations has also been analyzed.

Furthermore, an empirical comparison between the scale estimators $\widetilde{\mathrm{MAD}}$ and $\widetilde{\mathrm{AAD}}$ has been carried out and their breakdown point values have been indicated. The simulation study shows the robustness of $\widetilde{\mathrm{MAD}}$ in contrast to the non-robustness of $\widetilde{\mathrm{AAD}}$, when atypical values were introduced in the data.

In the future, it would be interesting to analyze consistency and other tools used to measure the robustness, such as the sensitive curve. In addition, alternatives to the robust scale measure $\widetilde{\mathrm{MAD}}$, like for instance the scale measure $\widetilde{\mathrm{IQR}}$, can be defined and studied for fuzzy data.

Acknowledgments. The research in this paper has been partially supported by the Severo Ochoa Grant BP12012 from the Principality of Asturias (de la Rosa de Sáa). Its financial support is gratefully acknowledged. De la Rosa de Sáa would like also thank the financial coverage through the Grant SV-PA-13-ECOEMP-66 for her short stay in the Vienna University of Technology supervised by Professor Peter Filzmoser.

References

1. Colubi, A.: Statistical Inference about the Means of Fuzzy Random Variables: Applications to the Analysis of Fuzzy- and Real-Valued Data. Fuzzy Sets Syst. 160(3), 344–356 (2009)
2. Diamond, P., Kloeden, P.: Metric spaces of fuzzy sets. Fuzzy Sets Syst. 100, 63–71 (1999)
3. Gil, M.A., Montenegro, M., González-Rodríguez, G., Colubi, A., Casals, M.R.: Bootstrap Approach to the Multi-Sample Test of Means with Imprecise Data. Comput. Stat. Data Anal. 51, 148–162 (2006)
4. González-Rodríguez, G., Colubi, A., Gil, M.A.: Fuzzy data treated as functional data. A one-way ANOVA test approach. Comput. Stat. Data Anal. 56(4), 943–955 (2012)
5. Hampel, F.R.: A general qualitative definition of robustness. Ann. Math. Statist. 42, 1887–1896 (1971)
6. Körner, R.: An asymptotic α-test for the expectation of random fuzzy variables. J. Stat. Plann. Inference 83, 331–346 (2000)
7. Lubiano, M.A., Gil, M.A., López-Díaz, M., López, M.T.: The $\boldsymbol{\lambda}$-Mean Squared Dispersion Associated with a Fuzzy Random Variable. Fuzzy Sets Syst. 111, 307–317 (2000)
8. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, New York (2006)
9. Montenegro, M., Colubi, A., Casals, M.R., Gil, M.A.: Asymptotic and Bootstrap Techniques for Testing the Expected Value of a Fuzzy Random Variable. Metrika 59, 31–49 (2004)

10. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
11. Ramos-Guajardo, A.B., Lubiano, M.A.: K-sample tests for equality of variances of random fuzzy sets. Compu. Stat. Data Anal. 56(4), 956–966 (2012)
12. Rousseeuw, P.J., Croux, C.: Alternatives to the Median Absolute Deviation. J. Am. Statist. Assoc. 88(424), 1273–1283 (1993)
13. Sinova, B., Gil, M.A., Colubi, A., Van Aelst, S.: The median of a random fuzzy number. The 1-norm distance approach. Fuzzy Sets Syst. 200, 99–115 (2012)
14. Vitale, R.A.: L_p metrics for compact, convex sets. J. Approx. Theor. 45, 280–287 (1985)
15. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning, Part 1. Inform. Sci. 8, 199–249 (1975); Part 2. Inform. Sci. 8, 301–353 (1975); Part 3. Inform. Sci. 9, 43–80 (1975)
16. Zieliński, R.: Robustness: a quantitative approach. Bull. Acad. Polon. Sci. Ser. Math. Astr. Phi. 24, 1281–1286 (1977)

The Wabl/Ldev/Rdev Median of a Random Fuzzy Number and Statistical Properties

Beatriz Sinova[1,2], Sonia Pérez-Fernández[1], and Manuel Montenegro[1]

[1] Departamento de Estadística e I.O. y D.M.,
Universidad de Oviedo, 33007 Oviedo, Spain
{sinovabeatriz,mmontenegro}@uniovi.es, soniapfdez@gmail.com
[2] Department of Applied Mathematics, Computer Science and Statistics,
Ghent University, 9000 Gent, Belgium

Abstract. In this paper the population and sample medians of a real-valued random variable are generalized to deal with random fuzzy numbers. The extension is based on a representation of fuzzy numbers for which necessary and sufficient conditions to characterize them have been previously established. Relevant statistical properties for these medians are studied.

Keywords: fuzzy data, robustness, wabl/ldev/rdev representation of a fuzzy number, wabl/ldev/rdev median of a random fuzzy number.

1 Introduction

In previous papers (see [3], [4]) extensions for the median notion to random fuzzy numbers (or RFN, in Puri and Ralescu's sense, see [1]) based on L^1 metrics have been introduced and discussed. The two metrics involved in those definitions make use of representations of fuzzy numbers for which necessary and sufficient conditions to characterize them are known. A third alternative will be introduced in this paper, as a generalization of the Hausdorff-type median for random intervals (Sinova *et al.* [2]). It will be shown that it fulfills convenient properties.

2 The φ-Wabl/Ldev/Rdev Median for an RFN

Let $\mathcal{F}_c(\mathbb{R})$ denote the class of bounded fuzzy numbers. The population and sample φ-wabl/ldev/rdev medians for random fuzzy numbers are defined as follows:

Definition 1. *Given a probability space $(\Omega, \mathcal{A}, P)$, an absolutely continuous probability measure φ on the measurable space $([0,1], \mathcal{B}_{[0,1]})$ with positive mass function on $(0,1)$, $\theta > 0$ and an associated RFN $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$, the* ***population φ-wabl/ldev/rdev median(s)*** *of $\mathcal{X}$ is (are) the fuzzy number(s)*

$$\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X}) = \arg\min_{\widetilde{U} \in \mathcal{F}_c(\mathbb{R})} E\left(\mathscr{D}_{\theta}^{\varphi}(\mathcal{X}, \widetilde{U})\right),$$

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_17

whenever these expectations exist, where

$$\mathscr{D}_\theta^\varphi(\widetilde{U},\widetilde{V}) = |\mathrm{wabl}^\varphi(\widetilde{U}) - \mathrm{wabl}^\varphi(\widetilde{V})|$$

$$+\frac{\theta}{2}\int_{[0,1]} |\mathrm{ldev}^\varphi_{\widetilde{U}}(\alpha) - \mathrm{ldev}^\varphi_{\widetilde{V}}(\alpha)|\, d\varphi(\alpha) + \frac{\theta}{2}\int_{[0,1]} |\mathrm{rdev}^\varphi_{\widetilde{U}}(\alpha) - \mathrm{rdev}^\varphi_{\widetilde{V}}(\alpha)|\, d\varphi(\alpha),$$

with

$$\mathrm{wabl}^\varphi(\widetilde{U}) = \int_{[0,1]} \frac{\inf \widetilde{U}_\alpha + \sup \widetilde{U}_\alpha}{2}\, d\varphi(\alpha),$$

$$\mathrm{ldev}^\varphi_{\widetilde{U}}(\alpha) = \mathrm{wabl}^\varphi(\widetilde{U}) - \inf \widetilde{U}_\alpha,\ \mathrm{rdev}^\varphi_{\widetilde{U}}(\alpha) = \sup \widetilde{U}_\alpha - \mathrm{wabl}^\varphi(\widetilde{U}).$$

Definition 2. *Given a probability space $(\Omega, \mathcal{A}, P)$, an absolutely continuous probability measure φ on the measurable space $([0,1], \mathcal{B}_{[0,1]})$ with positive mass function on $(0,1)$, $\theta > 0$, an associated RFN $\mathcal{X}$ and a simple random sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ obtained from $\mathcal{X}$, the* ***sample φ-wabl/ldev/rdev median(s)*** *of $\mathcal{X}$ is(are) the fuzzy number-valued statistic(s)*

$$\widehat{\widetilde{\mathrm{M}}^\varphi(\mathcal{X})}_n = \arg \min_{\widetilde{U}\in\mathcal{F}_c(\mathbb{R})} \frac{1}{n}\sum_{i=1}^n \left(\mathscr{D}_\theta^\varphi(\mathcal{X}_i, \widetilde{U})\right).$$

A key question at this stage is whether the φ-wabl/ldev/rdev median exists and can be computed easily in practice. The following result guarantees that at least one such median always exists and its computation is straightforward.

Theorem 1. *Given a probability space $(\Omega, \mathcal{A}, P)$, an absolutely continuous probability measure φ on the measurable space $([0,1], \mathcal{B}_{[0,1]})$ with positive mass function on $(0,1)$ and an associated RFN $\mathcal{X}$, for any $\alpha \in [0,1]$, the fuzzy number $\widetilde{\mathrm{M}}^\varphi(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$ such that*

$$\left(\widetilde{\mathrm{M}}^\varphi(\mathcal{X})\right)_\alpha = \left[\mathrm{Me}\big(\mathrm{wabl}^\varphi(\mathcal{X})\big) - \mathrm{Me}\big(\mathrm{ldev}^\varphi_{\mathcal{X}}(\alpha)\big), \mathrm{Me}\big(\mathrm{wabl}^\varphi(\mathcal{X})\big) + \mathrm{Me}\big(\mathrm{rdev}^\varphi_{\mathcal{X}}(\alpha)\big)\right]$$

(where in case $\mathrm{Me}\big(\mathrm{wabl}^\varphi(\mathcal{X})\big)$, $\mathrm{Me}\big(\mathrm{ldev}^\varphi_{\mathcal{X}}(\alpha)\big)$ *or* $\mathrm{Me}\big(\mathrm{rdev}^\varphi_{\mathcal{X}}(\alpha)\big)$ *are non-unique, the most usual convention for real-valued medians of choosing the midpoint of the interval of medians is considered) is a population φ-wabl/ldev/rdev median of $\mathcal{X}$.*

Proof. First, whatever $\alpha \in [0,1]$ and $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$ may be, we have that:

$$E\left[|\mathrm{wabl}^\varphi(\mathcal{X}) - \mathrm{Me}\big(\mathrm{wabl}^\varphi(\mathcal{X})\big)|\right] \le E\left[|\mathrm{wabl}^\varphi(\mathcal{X}) - \mathrm{wabl}^\varphi(\widetilde{U})|\right],$$
$$E\left[|\mathrm{ldev}^\varphi_{\mathcal{X}}(\alpha) - \mathrm{Me}\big(\mathrm{ldev}^\varphi_{\mathcal{X}}(\alpha)\big)|\right] \le E\left[|\mathrm{ldev}^\varphi_{\mathcal{X}}(\alpha) - \mathrm{ldev}^\varphi_{\widetilde{U}}(\alpha)|\right],$$
$$E\left[|\mathrm{rdev}^\varphi_{\mathcal{X}}(\alpha) - \mathrm{Me}\big(\mathrm{rdev}^\varphi_{\mathcal{X}}(\alpha)\big)|\right] \le E\left[|\mathrm{rdev}^\varphi_{\mathcal{X}}(\alpha) - \mathrm{rdev}^\varphi_{\widetilde{U}}(\alpha)|\right],$$

because $\mathrm{ldev}^\varphi_{\widetilde{U}}(\alpha)$, $\mathrm{rdev}^\varphi_{\widetilde{U}}(\alpha), \mathrm{wabl}^\varphi(\widetilde{U}) \in \mathbb{R}$, and $\mathrm{ldev}^\varphi_{\mathcal{X}}(\alpha)$, $\mathrm{rdev}^\varphi_{\mathcal{X}}(\alpha)$, and $\mathrm{wabl}^\varphi(\mathcal{X})$ are real-valued random variables. Therefore,

$$E\left(\mathscr{D}_\theta^\varphi(\mathcal{X}, \widetilde{U})\right) \ge E\left(\mathscr{D}_\theta^\varphi(\mathcal{X}, \widetilde{\mathrm{M}}^\varphi(\mathcal{X}))\right).$$

Now let's see that $\widetilde{\mathrm{M}}^\varphi(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$. Sufficient conditions to characterize fuzzy numbers are stated in Sinova *et al.* [5]:

- Over all Ω we have that $\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)$ and $\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha)$ are non-increasing functions of α in $[0,1]$, so due to the considered convention, $\mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))$ and $\mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))$ are non-increasing functions of α in $[0,1]$.
- Furthermore, functions $\mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))$ and $\mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))$ are left-continuous at every $\alpha \in (0,1]$:
  Indeed, if $\{\alpha_n\}_n \uparrow \alpha \in (0,1]$ as $n \to \infty$, then for any element in Ω we have that $\{\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n)\}_n \downarrow \mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)$ and, because of the considered convention, the sequence $\big\{\mathrm{Me}\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n)\big)\big\}_n \downarrow$ is bounded below. Since $\mathrm{Me}\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)\big)$ is a lower bound, there exists a limit for this sequence (it will be denoted by L^{φ}_{α}). Therefore, $L^{\varphi}_{\alpha} = \lim_{n\to\infty} \mathrm{Me}\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n)\big) \geq \mathrm{Me}\big(\mathrm{ldev}\, \mathcal{X}_{\alpha}\big)$, so:
  $$0.5 \leq P\Big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n) \geq \mathrm{Me}\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n)\big)\Big) \leq P\Big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n) \leq L^{\varphi}_{\alpha}\Big)$$
  for all $\omega \in \Omega$ using the definition of $\mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n))$. Since
  $$\Big\{\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n) \geq L^{\varphi}_{\alpha}\big)\Big\}_n \downarrow \bigcap_{n=1}^{\infty} \big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n) \geq L^{\varphi}_{\alpha}\big) = \big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha) \geq L^{\varphi}_{\alpha}\big),$$
  we have that
  $$P\Big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha) \geq L^{\varphi}_{\alpha}\Big) = P\Big(\lim_{n\to\infty} \big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha_n) \geq L^{\varphi}_{\alpha}\big)\Big) = \lim_{n\to\infty} P\Big(\mathrm{ldev}^{\varphi}_{\mathcal{X}} \geq L^{\varphi}_{\alpha}\Big) \geq 0.5.$$
  Following similar arguments, $P\Big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha) > L^{\varphi}_{\alpha}\Big) \leq 0.5$.
  Consequently, taking into account the considered convention, we have that $L^{\varphi}_{\alpha} = \mathrm{Me}\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)\big)$.
  Analogously, if $\{\alpha_n\}_n \uparrow \alpha \in (0,1]$ as $n \to \infty$, it holds that $\{\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha_n)\}_n \downarrow \mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha)$ and the sequence $\big\{\mathrm{Me}\big(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha_n)\big)\big\}_n \downarrow$, being bounded below by $\mathrm{Me}\big(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha)\big)$. Therefore, there exists
  $$L^{'\varphi}_{\alpha} = \lim_{n\to\infty} \mathrm{Me}\big(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha_n)\big)$$
  and following a reasoning like above, $L^{'\varphi}_{\alpha} = \mathrm{Me}\big(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha)\big)$.
- The right-continuity at 0 of both, $\mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))$ and $\mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))$, can be proved by means of similar arguments.
- Since $-\mathrm{ldev}^{\varphi}_{\mathcal{X}}(1) \leq \mathrm{rdev}^{\varphi}_{\mathcal{X}}(1)$ over all Ω, one can guarantee (using the considered convention) that:
  $$-\mathrm{ldev}^{\varphi}_{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}(1) = \mathrm{Me}(-\mathrm{ldev}^{\varphi}_{\mathcal{X}}(1)) \leq \mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(1)) = \mathrm{rdev}^{\varphi}_{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}(1).$$
- Finally,
  $$\int_{[0,1]} \mathrm{ldev}^{\varphi}_{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}(\alpha)\, d\varphi(\alpha) = \int_{[0,1]} \frac{\mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)) + \mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))}{2}\, d\varphi(\alpha)$$
  $$= \int_{[0,1]} \mathrm{rdev}^{\varphi}_{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}(\alpha)\, d\varphi(\alpha),$$
  whence $\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})$ is a bounded fuzzy number.

□

It is important to remark that the population φ-wabl/ldev/rdev median does not depend on the parameter θ, as it happened with the Hausdorff-type median for random intervals. The same remark would be also applicable to the sample φ-wabl/ldev/rdev median.

3 Statistical Properties of the φ-Wabl/Ldev/Rdev Median of an RFN

The φ-wabl/ldev/rdev median of a random fuzzy number preserves most of the basic properties of the median of a random variable. Thus, it can be straightforwardly proved that

Proposition 1. $\widetilde{\mathrm{M}}^{\varphi}$ *is equivariant under 'linear' transformations, that is, if* $\gamma \in \mathbb{R}$, $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$ *and* $\mathcal{X}$ *is an RFN, then*

$$\widetilde{\mathrm{M}}^{\varphi}(\gamma \cdot \mathcal{X} + \widetilde{U}) = \gamma \cdot \widetilde{\mathrm{M}}^{\varphi}(\mathcal{X}) + \widetilde{U},$$

where operations between fuzzy numbers are assumed to be based on Zadeh's extension principle. Consequently, if $\mathcal{X}$ *is a random fuzzy number associated with the probability space* $(\Omega, \mathcal{A}, P)$ *and the distribution of* $\mathcal{X}$ *is degenerate at a fuzzy number* $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$ *(i.e.,* $\mathcal{X} = \widetilde{U}$ *a.s.* $[P]$*), then* $\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X}) = \widetilde{U}$.

Proposition 2. *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space,* φ *be an absolutely continuous probability measure on the measurable space* $([0,1], \mathcal{B}_{[0,1]})$ *with positive mass function on* $(0,1)$ *and let* $\mathcal{X}$ *be a symmetric random fuzzy number about* $c \in \mathbb{R}$. *Then, the* φ*-wabl/ldev/rdev median of* $\mathcal{X}$ *is a symmetric fuzzy number about* c.

Theorem 2. *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space,* φ *be an absolutely continuous probability measure on the measurable space* $([0,1], \mathcal{B}_{[0,1]})$ *with positive mass function on* $(0,1)$ *and let* $\mathcal{X}$ *be a random fuzzy number associated with* $(\Omega, \mathcal{A}, P)$ *and satisfying that* $\mathrm{Me}(\mathrm{wabl}^{\varphi}(\mathcal{X}))$, $\mathrm{Me}\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)\big)$ *and* $\mathrm{Me}\big(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha)\big)$ *are actually unique (for each* $\alpha \in [0,1]$ *in case of the two last medians).*

If the two sequences of the real-valued sample medians $\big\{\widehat{\mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))}_n\big\}_n$ *and* $\big\{\widehat{\mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))}_n\big\}_n$ *as functions of* α *over* $[0,1]$ *are both uniformly integrable, then the estimator* $\widehat{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}_n$ *is strongly consistent in* $\mathscr{D}^{\varphi}_{\theta}$*-sense (and hence, in the sense of all the topologically equivalent metrics), i.e.*

$$\lim_{n\to\infty} \mathscr{D}^{\varphi}_{\theta}\left(\widehat{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}_n, \widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})\right) = 0 \quad a.s.\,[P].$$

Proof. Indeed,

$$P\left(\lim_{n\to\infty} \mathscr{D}^{\varphi}_{\theta}\left(\widehat{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}_n, \widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})\right) = 0\right)$$

$$= P\Bigg(\left(\lim_{n\to\infty} |\widehat{\mathrm{Me}(\mathrm{wabl}^{\varphi}(\mathcal{X}))}_n - \mathrm{Me}(\mathrm{wabl}^{\varphi}(\mathcal{X}))| = 0\right)$$

$$\bigcap \left(\lim_{n\to\infty} \int_{[0,1]} |\widehat{\mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))}_n - \mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))| = 0\right)$$

$$\bigcap \left(\lim_{n\to\infty} \int_{[0,1]} |\widehat{\mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))}_n - \mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))| = 0\right)\Bigg).$$

On one hand,

$$P\left(\lim_{n\to\infty} |\widehat{\mathrm{Me}(\mathrm{wabl}^{\varphi}(\mathcal{X}))}_n - \mathrm{Me}(\mathrm{wabl}^{\varphi}(\mathcal{X}))| = 0\right)$$

$$= P\left(\lim_{n\to\infty}\left(\mathrm{Me}(\widehat{\mathrm{wabl}^{\varphi}}(\mathcal{X}))_n - \mathrm{Me}(\mathrm{wabl}^{\varphi}(\mathcal{X}))\right) = 0\right) = 1,$$

due to the strong consistency of $\mathrm{Me}(\widehat{\mathrm{wabl}^{\varphi}}(\mathcal{X}))_n$.

On the other hand, under the assumption of uniqueness for the medians of $\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)$ and $\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha)$, the sample medians are strongly consistent estimators of the population medians, and hence

$$P\left(\lim_{n\to\infty}\left(\mathrm{Me}(\widehat{\mathrm{ldev}^{\varphi}_{\mathcal{X}}}(\alpha))_n - \mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))\right) = 0\right) = 1,$$
$$P\left(\lim_{n\to\infty}\left(\mathrm{Me}(\widehat{\mathrm{rdev}^{\varphi}_{\mathcal{X}}}(\alpha))_n - \mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))\right) = 0\right) = 1.$$

Moreover, assumptions for $\mathrm{Me}(\widehat{\mathrm{ldev}^{\varphi}_{\mathcal{X}}}(\alpha))_n$ and $\mathrm{Me}\big(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha)\big)$ guarantee that conditions to apply Vitali's Convergence Theorem are fulfilled, whence

$$P\left(\left(\lim_{n\to\infty}\int_{[0,1]}|\mathrm{Me}(\widehat{\mathrm{ldev}^{\varphi}_{\mathcal{X}}}(\alpha))_n - \mathrm{Me}(\mathrm{ldev}^{\varphi}_{\mathcal{X}}(\alpha))|\,d\varphi(\alpha) = 0\right)\right) = 1.$$

By following similar arguments, one can prove that

$$P\left(\left(\lim_{n\to\infty}\int_{[0,1]}|\mathrm{Me}(\widehat{\mathrm{rdev}^{\varphi}_{\mathcal{X}}}(\alpha))_n - \mathrm{Me}(\mathrm{rdev}^{\varphi}_{\mathcal{X}}(\alpha))|\,d\varphi(\alpha) = 0\right)\right) = 1.$$

Consequently,

$$P\left(\lim_{n\to\infty}\mathscr{D}^{\varphi}_{\theta}\left(\widehat{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}_n, \widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})\right) = 0\right). \qquad \square$$

The robustness of the φ-wabl/ldev/rdev median w.r.t. the mean will be now analyzed through the finite sample breakdown point of the sample median in a sample of size n from a random fuzzy number $\mathcal{X}$, which is now given by

$$\mathrm{fsbp}(\widehat{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}_n, \widetilde{\mathbf{x}}_n, \mathscr{D}^{\varphi}_{\theta})$$
$$= \frac{1}{n}\min\left\{k \in \{1,\ldots,n\} \,:\, \sup_{Q_{n,k}} \mathscr{D}^{\varphi}_{\theta}(\widehat{\widetilde{\mathrm{M}}^{\varphi}(P_n)}, \widehat{\widetilde{\mathrm{M}}^{\varphi}(Q_{n,k})}) = \infty\right\},$$

where $\widetilde{\mathbf{x}}_n$ denotes the considered sample of n data from the metric space $(\mathcal{F}_c(\mathbb{R}), \mathscr{D}^{\varphi}_{\theta})$ in which $\sup_{\widetilde{U},\widetilde{V}\in\mathcal{F}_c(\mathbb{R})} \mathscr{D}^{\varphi}_{\theta}(\widetilde{U},\widetilde{V}) = \infty$, P_n is the empirical distribution of $\widetilde{\mathbf{x}}_n$ and $Q_{n,k}$ is the empirical distribution of sample $\widetilde{\mathbf{y}}_{n,k}$ obtained from the original one $\widetilde{\mathbf{x}}_n$ by perturbing at most k components. Then, we have that

Proposition 3. *The finite sample breakdown point of the sample φ-wabl/ldev/rdev median from a random fuzzy number $\mathcal{X}$,* $\mathrm{fsbp}\big(\widehat{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}_n\big)$*, equals*

$$\mathrm{fsbp}\big(\widehat{\widetilde{\mathrm{M}}^{\varphi}(\mathcal{X})}_n\big) = \frac{1}{n}\cdot\lfloor\frac{n+1}{2}\rfloor,$$

where $\lfloor\cdot\rfloor$ denotes the floor function.

Proof. First note that the condition $\sup_{\widetilde{U},\widetilde{V}\in\mathcal{F}_c(\mathbb{R})} \mathscr{D}^{\varphi}_{\theta}(\widetilde{U},\widetilde{V}) = \infty$ is satisfied in this case, since $\mathscr{D}^{\varphi}_{\theta}\big(\mathbb{1}_{[n-1,n+1]}, \mathbb{1}_{[-n-1,-n+1]}\big) = 2n$ (of course, other examples could be provided for the same purpose).

Furthermore,

$$\mathscr{D}_\theta^\varphi(\widehat{\widetilde{\mathrm{M}}^\varphi(P_n)}, \widehat{\widetilde{\mathrm{M}}^\varphi(Q_{n,k})}) \geq |\mathrm{wabl}^\varphi(\widehat{\widetilde{\mathrm{M}}^\varphi(P_n)}) - \mathrm{wabl}^\varphi(\widehat{\widetilde{\mathrm{M}}^\varphi(Q_{n,k})})|$$

$$= |\mathrm{Me}(\widehat{\mathrm{wabl}^\varphi}(P_n)) - \mathrm{Me}(\widehat{\mathrm{wabl}^\varphi(Q_{n,k})})|.$$

Therefore, by recalling the fsbp of the sample median of a real-valued random variable, one can conclude that whenever at least $\lfloor\frac{n+1}{2}\rfloor$ elements $\widetilde{x}_i \in \mathcal{F}_c(\mathbb{R})$ of $\widetilde{\mathbf{x}}_n$ are replaced by other arbitrarily 'large' elements in $\mathcal{F}_c(\mathbb{R})$ so that

$$\sup_{Q_{n,k}} |\mathrm{Me}(\widehat{\mathrm{wabl}^\varphi}(P_n)) - \mathrm{Me}(\widehat{\mathrm{wabl}^\varphi(Q_{n,k})})| = \infty,$$

we have that

$$\sup_{Q_{n,k}} \mathscr{D}_\theta^\varphi(\widehat{\widetilde{\mathrm{M}}^\varphi(P_n)}, \widehat{\widetilde{\mathrm{M}}^\varphi(Q_{n,k})}) \geq \sup_{Q_{n,k}} |\mathrm{Me}(\widehat{\mathrm{wabl}^\varphi}(P_n)) - \mathrm{Me}(\widehat{\mathrm{wabl}^\varphi(Q_{n,k})})| = \infty,$$

whence

$$\mathrm{fsbp}(\widehat{\widetilde{\mathrm{M}}^\varphi(\mathcal{X})}_n, \widetilde{\mathbf{x}}_n, \mathscr{D}_\theta^\varphi) \leq \frac{1}{n} \cdot \lfloor\frac{n+1}{2}\rfloor.$$

On the other hand, by using again the fsbp of the sample median of a real-valued random variable, we have that for all $\alpha \in [0,1]$

$$\min\left\{k \in \{1,\dots,n\} \,:\, \sup_{Q_{n,k}} |\mathrm{Me}(\widehat{\mathrm{wabl}^\varphi}(P_n)) - \mathrm{Me}(\widehat{\mathrm{wabl}^\varphi(Q_{n,k})})| = \infty\right\} = \lfloor\frac{n+1}{2}\rfloor,$$
$$\min\left\{k \in \{1,\dots,n\} \,:\, \sup_{Q_{n,k}} |\mathrm{Me}\big(\widehat{\mathrm{ldev}_{P_n}^\varphi}(\alpha)\big) - \mathrm{Me}\big(\widehat{\mathrm{ldev}_{Q_{n,k}}^\varphi}(\alpha)\big)| = \infty\right\} = \lfloor\frac{n+1}{2}\rfloor,$$
$$\min\left\{k \in \{1,\dots,n\} \,:\, \sup_{Q_{n,k}} |\mathrm{Me}\big(\widehat{\mathrm{rdev}_{P_n}^\varphi}(\alpha)\big) - \mathrm{Me}\big(\widehat{\mathrm{rdev}_{Q_{n,k}}^\varphi}(\alpha)\big)| = \infty\right\} = \lfloor\frac{n+1}{2}\rfloor,$$

whence for all $\alpha \in [0,1]$

$$\sup_{Q_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\mathrm{wabl}^\varphi}(P_n)) - \mathrm{Me}(\widehat{\mathrm{wabl}^\varphi(Q_{n,k})})| = M_1 < \infty,$$
$$\sup_{Q_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}\big(\widehat{\mathrm{ldev}_{P_n}^\varphi}(\alpha)\big) - \mathrm{Me}\big(\widehat{\mathrm{ldev}_{Q_{n,k}}^\varphi}(\alpha)\big)| = M_2 < \infty,$$
$$\sup_{Q_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}\big(\widehat{\mathrm{rdev}_{P_n}^\varphi}(\alpha)\big) - \mathrm{Me}\big(\widehat{\mathrm{rdev}_{Q_{n,k}}^\varphi}(\alpha)\big)| = M_3 < \infty,$$

and therefore

$$\sup_{Q_{n,\lfloor\frac{n+1}{2}\rfloor-1}} \mathscr{D}_\theta^\varphi(\widehat{\widetilde{\mathrm{M}}^\varphi(P_n)}, \widehat{\widetilde{\mathrm{M}}^\varphi(Q_{n,\lfloor\frac{n+1}{2}\rfloor-1})})$$
$$= \sup_{Q_{n,\lfloor\frac{n+1}{2}\rfloor-1}} \Bigg[|\mathrm{Me}(\widehat{\mathrm{wabl}^\varphi}(P_n)) - \mathrm{Me}(\widehat{\mathrm{wabl}^\varphi(Q_{n,k})})|$$
$$+ \frac{1}{2}\int_{[0,1]} |\mathrm{Me}\big(\widehat{\mathrm{ldev}_{P_n}^\varphi}(\alpha)\big) - \mathrm{Me}\big(\widehat{\mathrm{ldev}_{Q_{n,\lfloor\frac{n+1}{2}\rfloor-1}}^\varphi}(\alpha)\big)|\, d\varphi(\alpha)$$
$$+ \frac{1}{2}\int_{[0,1]} |\mathrm{Me}\big(\widehat{\mathrm{ldev}_{P_n}^\varphi}(\alpha)\big) - \mathrm{Me}\big(\widehat{\mathrm{ldev}_{Q_{n,\lfloor\frac{n+1}{2}\rfloor-1}}^\varphi}(\alpha)\big)|\, d\varphi(\alpha)\Bigg]$$
$$\leq M_1 + \frac{M_2 + M_3}{2} < \infty.$$

Consequently,

$$\min\left\{k\in\{1,\ldots,n\} \,:\, \sup_{Q_{n,k}} \mathscr{D}_\theta^\varphi(\widehat{\widetilde{\mathrm{M}}^\varphi(P_n)}, \widehat{\widetilde{\mathrm{M}}^\varphi(Q_{n,k})}) = \infty\right\} > \lfloor\frac{n+1}{2}\rfloor - 1,$$

so that $\mathrm{fsbp}(\widehat{\widetilde{\mathrm{M}}^\varphi(\mathcal{X})}_n, \widetilde{\mathbf{x}}_n, \mathscr{D}_\theta^\varphi) \geq \frac{1}{n}\cdot\lfloor\frac{n+1}{2}\rfloor$. □

The following result formalizes the comparison of the robustness of the sample φ-wabl/ldev/rdev median and the sample mean of a random fuzzy number. Thus,

Proposition 4. *The finite sample breakdown point of the sample mean from a random fuzzy number $\mathcal{X}$, $\mathrm{fsbp}\big(\overline{\mathcal{X}_n}\big) = 1/n$, is lower than that for the sample φ-wabl/ldev/rdev median for sample sizes $n > 2$.*

The following simulations illustrate an empirical comparison between the mean, the 1-norm median (see [4]) and the ℓ-wabl/ldev/rdev median (where ℓ denotes the Lebesgue measure on $[0,1]$):

Step 1. A sample of size $n = 100000$ of trapezoidal fuzzy numbers has been simulated for each of some different situations in such a way that

- to generate the trapezoidal fuzzy data, we have considered four real-valued random variables as follows: $X_1 = \mathrm{mid}\,\mathcal{X}_1$, $X_2 = \mathrm{spr}\,\mathcal{X}_1$, $X_3 = \inf\mathcal{X}_1 - \inf\mathcal{X}_0$, $X_4 = \sup\mathcal{X}_0 - \sup\mathcal{X}_1$;
- each sample is assumed to be split into a subsample of size $n(1-c_p)$ (c_p = proportion of contamination ranging in $\{0, 0.1, 0.2, 0.4\}$) associated with a non-contaminated distribution and a subsample of size $n\cdot c_p$ associated with a contaminated one, where an additional contamination role is played by C_D (which measures how far the distribution of the contaminated subsample is from the distribution of the non-contaminated one and ranges in $\{0, 1, 5, 10, 100\}$);
- 16 situations for different values of c_p and C_D have been considered for simulations and for each of these situations two cases have been selected, namely, one in which random variables X_i are independent (CASE 1) and another one in which they are dependent (CASE 2). More specifically, CASE 1 assumes that
  - •• $X_1 \sim \mathcal{N}(0,1)$ and $X_2, X_3, X_4 \sim \chi_1^2$ for the non-contaminated subsample,
  - •• $X_1 \sim \mathcal{N}(0,3) + C_D$ and $X_2, X_3, X_4 \sim \chi_4^2 + C_D$ for the contaminated subsample,

  whereas CASE 2 assumes that
  - •• $X_1 \sim \mathcal{N}(0,1)$ and $X_2, X_3, X_4 \sim 1/(X_1^2+1)^2 + 0.1\cdot\chi_1^2$ for the non-contaminated subsample (with χ_1^2 independent of X_1),
  - •• $X_1 \sim \mathcal{N}(0,3) + C_D$ and $X_2, X_3, X_4 \sim 1/(X_1^2+1)^2 + 0.1\cdot\chi_1^2 + C_D$ for the contaminated subsample (with χ_1^2 independent of X_1).

Step 2. $N = 1000$ replications of *Step 1* in the first simulations have been considered, so that for each of the 16 situations concerning c_p and c_D there are 1000 available samples of size $n = 100000$.

Step 3. For each of the 16 situations and 1000 replications, the mean distance between the non-contaminated distribution and each sample mixed location measure is computed. Finally, the mean over the 1000 samples (MD) is obtained.

Distances have been computed by considering the well-known ρ_2. The outputs for this simulation study have been collected in Table 1.

Simulations support empirically the fact that the contamination affects the Aumann-type mean much more than the two considered medians. Moreover, the wabl/ldev/rdev median behaves in a slightly more robust way than the 1-norm one in both CASES 1 and 2.

Table 1. Mean ρ_2-distance of the location measure to the non-contaminated distribution

		CASE 1			CASE 2		
c_P	c_D	Aumann mean	1-norm median	w/l/r median	Aumann mean	1-norm median	w/l/r median
0	0	1.590684	1.552950	1.554372	1.002750	1.032440	1.009203
0.1	0	1.685077	1.564486	1.553082	1.004752	1.031691	1.012825
0.1	1	1.727329	1.569681	1.553262	0.990237	1.037552	0.994915
0.1	5	1.958604	1.566279	1.554583	1.085708	1.043102	0.996064
0.1	10	2.355203	1.568843	1.554976	1.568303	1.044393	0.996751
0.1	100	13.552122	1.569227	1.555065	13.187656	1.045329	0.996780
0.2	0	1.825401	1.593075	1.563680	1.007136	1.030994	1.016413
0.2	1	1.946608	1.602914	1.565674	0.988365	1.043201	0.986686
0.2	5	2.601743	1.615051	1.572603	1.548887	1.056944	0.992840
0.2	10	3.658885	1.617811	1.574932	2.728677	1.058855	0.994872
0.2	100	26.827333	1.617947	1.575027	26.308354	1.061132	0.993515
0.4	0	2.212263	1.759157	1.711233	1.012141	1.028484	1.026757
0.4	1	2.558034	1.865478	1.742423	1.025340	1.061503	1.015477
0.4	5	4.194068	2.014625	1.802162	2.701750	1.124580	1.067675
0.4	10	6.646680	2.092274	1.817192	5.211615	1.138088	1.091136
0.4	100	54.186654	2.101532	1.829062	51.864586	1.142909	1.091821

4 Concluding Remarks

A new approach for the median of a random fuzzy number has been introduced. Some of its properties have been proved and its robustness has been shown by calculating its finite sample breakdown point, empirically checked through some simulations. Among the future directions, the sensitivity analysis of the influence of φ on the resulting median will be the main aspect to bear in mind.

Acknowledgments. The research in this paper has been partially supported by the FPU Grant AP2009-1197 (Sinova) and a Beca de colaboración 2013-2014 (Pérez-Fernández) from the Spanish Ministry of Education. The support is gratefully acknowledged.

References

1. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
2. Sinova, B., Casals, M.R., Colubi, A., Gil, M.Á.: The median of a random interval. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano, M.A., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. AISC, vol. 77, pp. 575–583. Springer, Heidelberg (2010)
3. Sinova, B., De la Rosa de Sáa, S., Gil, M.A.: A generalized L^1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data. Inform. Sci. 242, 22–34 (2013)
4. Sinova, B., Gil, M.A., Colubi, A., Van Aelst, S.: The median of a random fuzzy number. The 1-norm distance approach. Fuzzy Sets Syst. 200, 99–115 (2012)
5. Sinova, B., Gil, M.A., López, M.T., Van Aelst, S.: A parameterized L^2 metric between fuzzy numbers and its parameter interpretation. Fuzzy Sets Syst. 245, 101–115 (2014)

Chi-Square Test for Homogeneity with Fuzzy Data

Przemysław Grzegorzewski[1,2] and Hubert Szymanowski[3]

[1] Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw, Poland
[2] Faculty of Mathematics and Information Science,
Warsaw University of Technology,
Koszykowa 75, 00-662 Warsaw, Poland
[3] Institute of Computer Science, Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warsaw, Poland
pgrzeg@ibspan.waw.pl,
h.szymanowski@ipipan.waw.pl

Abstract. Fuzzy opinions are very common in surveys performed by social sciences. A fuzzy multinomial distribution for modeling such opinions is proposed. Next, a method for constructing a generalized version of the chi-square test of homogeneity which allows fuzzy data is proposed.

Keywords: categorical data, chi-square test of homogeneity, fuzzy answer, fuzzy data, multinomial distribution, preference, questionnaires.

1 Introduction

A test of homogeneity involves testing that the proportions of elements with certain category in two or more populations are the same. More formally, testing homogeneity means verification whether several multinomial distributions corresponding to particular populations are similar (homogeneous).

The best known test of homogeneity is the chi-square test applied to sample data organized in a contingency table. The data often come from questionnaires, popular especially in social sciences. Traditionally, the respondent examined during a survey should choose his/her favorite category from a list of given options. Usually these options are mutually exclusive and exhaustive, i.e. the respondent should indicate one and only one option.

The classical chi-square test of homogeneity requires contingency tables with exclusive categories. The last assumptions often appears too rigid in practice, because the respondents often can hardly choose their favorite options. Thus fuzzy answers allowing to specify a degree of conviction for each category, to which it is the most preferred one, seems to be useful. Moreover, a generalization of the chi-square test of homogeneity allowing fuzzy answers would be desirable.

Formally the chi-square test of homogeneity is a goodness-of-fit test for the multinomial distribution. Although many statistical tools have been generalized for fuzzy data, there are only a few papers devoted to goodness-of-fit tests in a fuzzy environment [2–6]. The goal of this paper is to generalize not only the

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_18

test but also the notion of the multinomial distribution. In this aspect our idea is close to that proposed by Lin et al. [5], however we propose a model which - in our opinion - suits better to the nature of fuzzy answers in questionnaires.

The paper is organized as follows: In Sec. 2 we propose a way for modeling fuzzy answers leading to fuzzy multinomial distribution. Next, in Sec. 3 we show how to construct the chi-square test of homogeneity in fuzzy environment. All suggested notions and tools are illustrated by examples.

2 Fuzzy Preferences and Their Distributions

2.1 Fuzzy Answers in Questionnaires

Suppose that a given question in a survey admits d options. Usually the answer to that question might be identified with $\mathbb{X} = (X_1, \ldots, X_d)$, where $X_i \in \{0, 1\}$ such that $\sum_i^d X_i = 1$. Here $X_i = 1$ indicates that one chooses the i-th option.

In many cases exclusive choices are not natural and too restrictive and it seems that fuzzy answers would be much more appropriate there. In other words, instead of choosing one and only one category (option), the respondent may divide his/her vote among several options proportionally to his/her convictions or preferences. Therefore, a *fuzzy answer* might be identified with a vector $\mathbb{X} = (X_1, \ldots, X_d)$, where $\sum_i^d X_i = 1$, but now $X_i \in [0, 1]$. In this case X_i indicates the grade of preference attributed to i-th option. If someone attributes $X_i = 1$ to i-th category, it would be interpreted that he/she is completely convinced to this option. Such a *crisp answer* is, of course, a particular case of a fuzzy answer.

2.2 Fuzzy Multinomial Distribution

Having two categories $(X_1, X_2) = (X_1, 1 - X_1)$ one may identify the first option with *success* if $X_1 \geq 0.5$. Similarly, for $d \geq 2$ the label *success* will be attributed to the option with the highest grade of preference. Let us denote the probability that the i-th option is classified as *success* by

$$\pi_i = P(X_i = \max\{X_1, ..., X_d\}). \tag{1}$$

Before we generalize the multinomial distribution so it could be applied as an adequate mathematical model for fuzzy answers, we need some assumptions on the X_i distribution. Since there is no reason neither to favor nor to discriminate any option, some kind of averaging seems to be reasonable. Therefore, we assume that the distribution of $X_i | (X_i = \max\{X_1, X_2, ..., X_d\})$ is uniform.

Definition 1. *Let S_d denote the d-dimensional simplex, i.e.*

$$S_d = \{(x_1, \ldots, x_d) \in \mathbb{R}^d : x_1, \ldots, x_d \geq 0, x_1 + \cdots + x_d = 1\} \tag{2}$$

and let

$$M_i = \{(x_1, \ldots, x_d) \in S_d : x_i = \max\{x_1, \ldots, x_d\}\}, \quad i = 1, \ldots, d. \tag{3}$$

*We say that $\mathbb{X} = (X_1, \ldots, X_d)$ has the **d-dimensional fuzzy multinomial distribution**, and we denote it as $\mathbb{X} \sim FM(d, \Pi)$, if its density is given by*

$$f(x_1, \ldots, x_d) = \begin{cases} d! \cdot \pi_1 & \textit{if} \quad (x_1, \ldots, x_d) \in M_1 \\ \vdots & \qquad \vdots \\ d! \cdot \pi_d & \textit{if} \quad (x_1, \ldots, x_d) \in M_d, \end{cases} \tag{4}$$

where $\Pi = (\pi_1, \ldots, \pi_d)$ such that $\pi_1, \ldots, \pi_d \geq 0$ and $\sum_{i=1}^{d} \pi_i = 1$.

One can prove that (4) is a probability distribution. Actually, since $\sum_{i=1}^{d} x_i = 1$, we may express any variable by the remaining ones. Let us fix x_d. Hence (4) is a function of $d-1$ independent variables and is positive on a set

$$\tilde{S}_{d-1} = \{(x_1, ..., x_{d-1}) \in \mathbb{R}^{d-1} : x_1, ..., x_{d-1} \geq 0, x_1 + \cdots + x_{d-1} \leq 1\},$$

whose Lebesgue measure is $\frac{1}{(d-1)!}$. By the symmetry the Lebesgue measures of each set M_i, $i = 1, \ldots, d$ are identical and equal to $\frac{1}{d!}$. Thus

$$\int_{\tilde{S}_{d-1}} f(x_1 \ldots, x_{d-1}) dx_1 \ldots dx_{d-1} = d! \left(\pi_1 \cdot \frac{1}{d!} + \ldots + \pi_d \cdot \frac{1}{d!} \right) = 1.$$

2.3 Some Examples

Let us consider two particular examples of the suggested fuzzy multinomial distribution.

Example 1. Let $\mathbb{X} = (X_1, X_2, X_3) \sim FM(3, \Pi)$, $\Pi = (\pi_1, \pi_2, \pi_3)$, which might be considered as a model of a fuzzy answer to a question admitting three options. By (4) we get

$$f(x, y, z) = \begin{cases} 6\pi_1 & \text{if} \quad (x, y, z) \in M_1 \\ 6\pi_2 & \text{if} \quad (x, y, z) \in M_2 \\ 6\pi_3 & \text{if} \quad (x, y, z) \in M_3 \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

where $\pi_1, \pi_2, \pi_3 \geq 0$ and $\pi_1 + \pi_2 + \pi_3 = 1$. Since $z = 1 - x - y$, substituting this relation into (5) we finally get the following density of a fuzzy multinomial distribution:

$$f(x, y) = \begin{cases} 6\pi_1 & \text{if} \quad y \geq 0, \ y \geq 1 - 2x, \ x \leq 1 - y, \ y \leq x \\ 6\pi_2 & \text{if} \quad x \geq 0, \ y > x, \ y \geq \frac{1-x}{2}, \ y \leq 1 - x \\ 6\pi_3 & \text{if} \quad x \geq 0, \ y \geq 0, \ y < \frac{1-x}{2}, \ y < 1 - 2x \\ 0 & \text{otherwise}. \end{cases} \tag{6}$$

Fig. 1 shows the subsets of the unit square where (6) is positive and a value assumed by $f(x, y)$ in each subset.

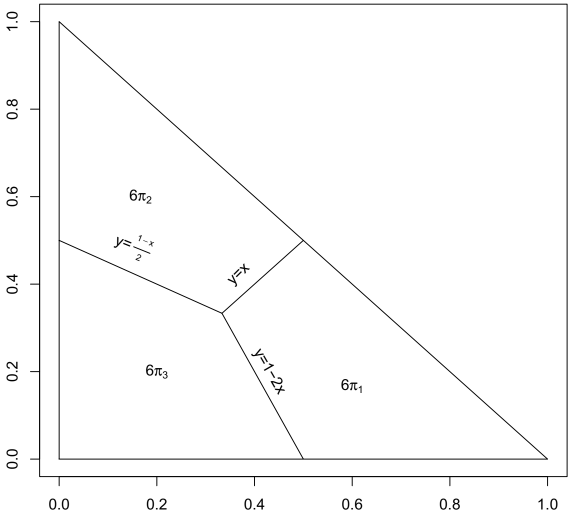


Fig. 1. Support and values of the $FM(3, \Pi)$ density

After long and tedious calculations we may show that the marginal density of each X_i $(i = 1, 2, 3)$ is given by

$$f(x) = \begin{cases} 3(1-x)(1-\pi_i) & \text{if} \quad x \in [0, \frac{1}{3}) \\ 6(1-2x) + 6\pi_i(5x-2) & \text{if} \quad x \in [\frac{1}{3}, \frac{1}{2}) \\ 6\pi_i(1-x) & \text{if} \quad x \in [\frac{1}{2}, 1] \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and hence its expected value and variance are equal to

$$\mu_i = \mathbb{E}X_i = \frac{7 + 15\pi_i}{36} \tag{8}$$

$$\sigma^2{}_i = Var(X_i) = \frac{20 + 231\pi_i - 225\pi_i^2}{1296}. \tag{9}$$

The covariance between any two marginals X_i, X_j, for $i \neq j$ is given by

$$Cov(X_i, X_j) = -\frac{7 + 6\pi_i + 6\pi_j + 225\pi_i\pi_j}{1296}. \tag{10}$$

□

Example 2. In the two-dimensional case, i.e. for $(X_1, X_2) \sim FM(2, \Pi)$, $\Pi = (\pi_1, \pi_2)$, which might be considered as a model of a fuzzy answer to the question admitting possible two options, the situation simplifies a lot. Actually, since now we have $(X, Y) = (X, 1 - X)$ and $\pi_2 = 1 - \pi_1$, therefore, by (4), we get

$$f(x) = \begin{cases} 2\pi_1 & \text{if} \quad x \in [0.5, 1] \\ 2(1-\pi_1) & \text{if} \quad x \in [0, 0.5). \end{cases} \tag{11}$$

It is worth noticing that (11) coincides with the fuzzy Bernoulli distribution proposed by Lin et al. [5], denoted as $X \sim FB(\pi_1)$. □

3 Testing Homogeneity with Fuzzy Data

3.1 Chi-square Test for Homogeneity

Suppose we like to test homogeneity of $k \geq 2$ populations with regard to the distribution of their $d \geq 2$ categories. Hence the distribution of the i-th population is represented by the multinomial distribution $\Pi_i = (\pi_{i1}, \ldots, \pi_{id})$. We verify the null hypothesis

$$H_0 : \Pi_1 = \cdots = \Pi_k = \Pi_0 = (\pi_{01}, \ldots, \pi_{0d}) \tag{12}$$

stating that there are no significant differences between distributions, against the alternative hypothesis $H_1 : \neg H_0$, that at least two distributions differ.

The most famous statistical tool for testing (12) is the chi-square test of homogeneity. It requires data organized in a contingency table given as follows:

	category 1	category 2	...	category d	Σ
population 1	O_{11}	O_{12}	...	O_{1d}	n_1
population 2	O_{21}	O_{22}	...	O_{2d}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
population k	O_{k1}	O_{k2}	...	O_{kd}	n_k
Σ	$O_{\cdot 1}$	$O_{\cdot 2}$	...	$O_{\cdot d}$	N

Here O_{ij} denotes the observed frequency of the j-th category in a sample from the i-th population, while n_i stands for the number of observations in i-th sample. We assume that the total number of observations in all k samples are $N = n_1 + \ldots + n_d$. Moreover, $O_{\cdot 1}, \ldots O_{\cdot d}$ are column totals.

To perform a test we compare the observed frequencies O_{ij} with the corresponding expected frequencies E_{ij}, where $E_{ij} = \frac{1}{N} n_i O_{\cdot j}$. If the null hypothesis (12) holds then the test statistic, given by

$$T = \sum_{i=1}^{k} \sum_{j=1}^{d} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{13}$$

is chi-square distributed with $(k-1)(d-1)$ degrees of freedom. Hence we reject H if the value of test statistic T is too large.

In the next section we show how to generalize the chi-square test of homogeneity for fuzzy multinomial distributions. Unfortunately, because of quite complicated calculations, this generalization has to be performed separately for each number of categories. Therefore, we have decided do propose below only how to construct the chi-square test of homogeneity for three categories.

3.2 Chi-square Test for Three Categories

Suppose that we want to verify whether the answers coming from k groups of respondents are homogeneous. In this section we assume that instead of exclusive choices between A, B and C we admit fuzzy answers, where each respondent specifies his/her grades of preferences between these three categories. In such a case we may gather all aggregated answers in the following table

	A	B	C	Σ
G_1	Z_{11}	Z_{12}	Z_{13}	n_1
G_2	Z_{21}	Z_{22}	Z_{23}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G_k	Z_{k1}	Z_{k2}	Z_{k3}	n_k
Σ	$Z_{\cdot 1}$	$Z_{\cdot 2}$	$Z_{\cdot 3}$	N

For each group of respondents $G_1, \ldots, G_k$ its aggregated answers might be perceived as a random variable $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$, such that

$$Z_{ij} = \sum_{l=1}^{n_i} X_{ijl} \quad j = 1, 2, 3, \tag{14}$$

where X_{ijl} denotes a grade attributed to the j-th category by the l-th respondent from the i-th group, n_i stands for the frequency of the i-th group, and

$$Z_{\cdot j} = \sum_{i=1}^{k} Z_{ij}, \tag{15}$$

Moreover, let $X_{il} = (X_{i1l}, X_{i2l}, X_{i3l})$ denote a vector corresponding to a fuzzy answer of the l-th respondent from the i-th group. Let us assume that $X_{i1}, \ldots, X_{in_i}$ are random vectors coming from the $FM(3, \pi_i)$ distribution, where $\Pi_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})$.

Our goal is to verify the null hypothesis

$$H_0 : \Pi_1 = \cdots = \Pi_k = \Pi_0 = (\pi_{01}, \pi_{02}, \pi_{03}) \tag{16}$$

stating that the answers in all groups are concordant, against the alternative hypothesis $H_1 : \neg H_0$, that the answers in at least two groups differ significantly.

Assuming that the null hypothesis holds, each X_{il} has the same distribution $FM(3, \Pi_0)$. Let $\mu = (\mu_1, \mu_2, \mu_3)$ denote its mean, given by (8). It seems that a reasonable estimator of $\mathbb{E}Z_{ij}$, $i = 1, \ldots, k$, $j = 1, 2, 3$, is

$$E_{ij} = n_i \cdot \hat{\mu}_j, \tag{17}$$

where

$$\hat{\mu}_j = \frac{Z_{\cdot j}}{N}. \tag{18}$$

Since $\mathbb{E}Z_{ij} = n_i\frac{7+15\pi_j}{36}$, the consistent estimator of π_j is given by

$$\hat{\pi}_j = \frac{36E_{ij}}{15n_i} - \frac{7}{15} = \frac{36Z_j}{7N} - \frac{7}{15} = \frac{36}{7}\hat{\mu}_j - \frac{7}{15}. \tag{19}$$

Some facts given in [1], a multivariate version of the Central Limit Theorem and the Slucky theorem [8] are useful to prove the following theorem.

Theorem 1. *Let*

$$W_1 = \frac{4(7+15\hat{\pi}_{01})(7+6\hat{\pi}_{02}+6\hat{\pi}_{03}+225\hat{\pi}_{02}\hat{\pi}_{03})}{3(13+168(\hat{\pi}_{01}\hat{\pi}_{02}+\hat{\pi}_{01}\hat{\pi}_{03}+\hat{\pi}_{02}\hat{\pi}_{03})+2025\hat{\pi}_{01}\hat{\pi}_{02}\hat{\pi}_{03})}, \tag{20}$$

$$W_2 = \frac{4(7+15\hat{\pi}_{02})(7+6\hat{\pi}_{01}+6\hat{\pi}_{03}+225\hat{\pi}_{01}\hat{\pi}_{03})}{3(13+168(\hat{\pi}_{01}\hat{\pi}_{02}+\hat{\pi}_{01}\hat{\pi}_{03}+\hat{\pi}_{02}\hat{\pi}_{03})+2025\hat{\pi}_{01}\hat{\pi}_{02}\hat{\pi}_{03})}, \tag{21}$$

$$W_3 = \frac{4(7+15\hat{\pi}_{03})(7+6\hat{\pi}_{01}+6\hat{\pi}_{02}+225\hat{\pi}_{01}\hat{\pi}_{02})}{3(13+168(\hat{\pi}_{01}\hat{\pi}_{02}+\hat{\pi}_{01}\hat{\pi}_{03}+\hat{\pi}_{02}\hat{\pi}_{03})+2025\hat{\pi}_{01}\hat{\pi}_{02}\hat{\pi}_{03})}. \tag{22}$$

Then, assuming the null hypothesis (16) holds, the following statistic

$$T_3 = \sum_{j=1}^{3} W_j \sum_{i=1}^{k} \frac{(Z_{ij}-E_{ij})^2}{E_{ij}}. \tag{23}$$

is asymptotically chi-square distributed with $2(k-1)$ *degrees of freedom.*

Test statistic (23) is very similar to the statistic of the classical chi-square test. Indeed, as in (13) we may distinguish observed and expected frequencies calculated for each cell of the contingency table. However, test statistic (23), contrary to its classical prototype, might be perceived as a weighted chi-square statistic with weights W_1, W_2 and W_3 given by (20)-(22). And, similarly as using the classical test, we reject the null hypothesis (16) if T_3 is too large, i.e. if T_3 exceeds the $(1-\alpha)100\%$ quantile of the chi-square distribution $\chi^2(2(k-1))$, where α is a significance level. Otherwise, one may compute an adequate p-value.

Example 3. Suppose we are interested whether there are any significant differences in preferences for favorite fruits between male and female. We asked 50 women and 50 men to choose their favorite fruits among bananas, apples and grapes. Contrary to the classical survey with exclusive choices each respondent could divide his/her vote between available fruits. So we got fuzzy answers like

$$\text{my favorite fruit} = b/\text{banana} + a/\text{apple} + g/\text{grape},$$

where $b, a, g \in [0,1]$, satisfying $b+a+g=1$, expressed the weights corresponding to each fruits considered as the most favorite one. All received answers after appropriate aggregation were organized in the following contingency table:

	Banana	Apple	Grape	Σ
Female	15	21.6	13.4	50
Male	13.5	19.2	17.3	50
Σ	28.5	40.8	30.7	100

Our goal is to verify the null hypothesis $H_0 : \Pi_F = \Pi_M = \Pi_0 = (\pi_{01}, \pi_{02}, \pi_{03})$, stating that there are no significant differences in preferences of male and female.

The firs step is to estimate E_{ij}. In our case $n_1 = n_2 = 50$, $N = 100$, while $Z_{\cdot 1} = 28.5$, $Z_{\cdot 1} = 40.8$ and $Z_{\cdot 1} = 30.7$ are obtained from our contingency table. By (17) and (18) and taking Z_{ij} form the contingency table we compute E_{ij}. As a result we get $E_{11} = E_{21} = 14.25$, $E_{12} = E_{22} = 20.4$ and $E_{13} = E_{23} = 15.35$. Now, by (18) and (19) we get: $\hat{\pi}_{01} = 0.217$, $\hat{\pi}_{02} = 0.513$ and $\hat{\pi}_{03} = 0.27$.

Substituting all those values into (20)-(22) and then into (23) we get $T_3 = 2.99$. The corresponding p-value is equal to 0.224 so there is no reason to reject H_0. It means that both male and female do not differ significantly in their preferences for the selected group of fruits. □

4 Conclusions

Fuzzy answers are natural and common in surveys considered in the social sciences. In this paper we proposed a methodology for modeling probability distributions corresponding to such data. Moreover, we showed how to construct a generalized version of the chi-square test of homogeneity which allows nonexclusive categories in contingency tables.

The main drawback of the generalized chi-square construction is that the shape of the test statistic depends on the number of the considered categories. However, since nowadays most of the calculations in statistics are perform with a professional software, this disadvantage could be practically diminished by using a suitable package (e.g. in R [7]).

References

1. Chernoff, H., Lehmann, E.L.: The use of maximum likelihood estimates in χ^2 tests for goodness of fit. Ann. Math. Statist. 25, 579–586 (1954)
2. Grzegorzewski, P., Jedrej, A.: Chi-square goodness-of-fit test for vague data. In: Proceedings of the Eleventh International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems - IPMU 2006, Paris, France, pp. 952–956 (2006)
3. Grzegorzewski, P., Szymanowski, H.: Goodness-of-fit tests for fuzzy data (submitted, 2014)
4. Hesamian, G., Taheri, S.M.: Fuzzy empirical distribution function: properties and application. Kybernetika 49, 962–982 (2013)
5. Lin, P.C., Wu, B., Watada, J.: Goodness-of-fit test for membership functions. International Journal of Innovative Computing, Information and Control 8, 7437–7450 (2012)
6. Nguyen, H., Wu, B.: Fundamentals of Statistics with Fuzzy Data. Springer, Heidelberg (2006)
7. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2013), `http://www.R-project.org/`
8. Slutsky, E.: Über stochastische Asymptoten und Grenzwerte. Metron 5, 3–89 (1925)

On Comparison of Distorted Histograms

Alexander Lepskiy[*]

Higher School of Economics,
20 Myasnitskaya Ulitsa, Moscow, 101000, Russia

Abstract. There are many tasks where the comparison of histograms (distributions, fuzzy numbers) is required with help of relationship of type "more-less". There are many approaches to solving this problem. But the histograms may be distorted. Then we have to find the conditions on the distortions under which the comparison of the two histograms is not changed. The solution of the problem is searched via three popular probabilistic methods of comparison.

Keywords: comparison of distributions, distortions of distributions, stability of comparison.

1 Introduction

There are many tasks where the comparison of histograms (distributions, fuzzy numbers) is necessary. We will consider the discrete distributions which are given with help of histograms of type $U = (x_i, u_i)_{i\in I}$, $x_i < x_{i+1}$, $i \in I$. Such histograms may consist of fuzzy discrete numbers. In this paper we will consider the comparison of type "more-less". For example, comparison of results of different experiments (see, e.g. [1]); comparison of functional indicators of the organizational, technical systems etc. [2]; decision-making under fuzzy uncertainty [3]; simulation of fuzzy preferences [4]; comparisons of income distribution within the framework of socio-economic analysis [5]; ranking students based on the results of histograms of their grades [6] etc. The different approaches are used for comparing histograms. Probabilistic approach is one of the most popular. Some numerical characteristics of random variables associated with histograms are compared in this approach. Another approach is based on the use of ranking methods of income distribution in the theory of social choice [5]. Histograms income has the form $U = (i, u_i)_{i=1}^{n_U} = (u_i)_{i=1}^{n_U}$, where $u_1 \leq u_2 \leq ... \leq u_{n_U}$ in this case. These histograms are compared with help of welfare functions $W(U)$ that satisfy the conditions of symmetry, monotonicity, concavity, etc. This approach is equivalent to ranking of ordered ascending vectors if the dimensions of vector-histograms are the same. The methods of the importance of criteria can be used in this case [7], or social threshold aggregations [8], etc.

[*] The study was implemented in the framework of The Basic Research Program of the Higher School of Economics. This work was supported by the grant 14-07-00189 of RFBR (Russian Foundation for Basic Research).

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_19

The third approach to ranking histograms is associated with the use of the tools of comparison of fuzzy numbers. The histogram $U = (x_i, u_i)_{i \in I}$ is associated with fuzzy set (or fuzzy number) [9] by means of membership function $U = (u_i)_{i \in I}$ which is defined on the universal set $X = (x_i)_{i \in I}$. Methods of comparison of fuzzy numbers can be used in this case [10,11,12]. Overview and analysis of the main approaches of comparing histograms are given in [6].

The comparison of histograms can be defined with some degree of imprecision. The nature of these imprecision may be different. For example, the uncertainty can be probabilistic character when compared histograms as the results of experiments. The imprecision may be the result of deliberate distortion of data in the theory of collective choice. The filling gap in incomplete data is another type of distortion.

Thus we have following questions. Can a distortion change the comparison of histograms by definite method to the opposite? Which distortion does not change the result of the comparison? The purpose of this paper is to obtain answers to these questions. In this paper we will analyze the stability to the distortion of some of the most popular probabilistic methods of comparing histograms.

2 Notation and Definitions

A pair $U = (x_i, u_i)_{i \in I}$ of two ordered sets of numbers will be treated as the histogram in this work, where $(x_i)_{i \in I}$ is an ordered ascending vector different arguments of histogram (i.e. $x_i < x_{i+1}$, $i \in I$), $(u_i)_{i \in I}$ is a vector of non-negative values of histogram, I is a some index set.

We are to define the total preorder relation R (reflexive, complete and transitive relation) on the set of histograms $\mathcal{U} = \{U\}$. If histograms U and V are in the relation R (i.e. $(U, V) \in R$)), then we will denote this through $U \succ V$ and we will define that U is greater than V. If $U \succ V$ and $V \succ U$ then we will treat these histograms as equal ad we will denote $U \sim V$.

We will also assume that the relation R should be in accordance with the condition, that ordering arguments of the histogram ascend their importance: if $U' = (x_i, u_i')$, $U'' = (x_i, u_i'')$ are two histograms for which $u_i' = u_i''$ for all $i \neq k, l$ and $u_l' - u_l'' = u_k'' - u_k' \geq 0$ then $U'' \succ U'$ for $k > l$ and $U' \succ U''$ for $k < l$.

Without loss of generality we can assume that the compared histograms are "aligned on the number of columns", i.e. if $U = (x_i^U, u_i)_{i \in I_U}$ and $V = (x_i^V, u_i)_{i \in I_V}$ are two histograms, then $I_U = I_V$ and $\{x_i^U\}_{i \in I} = \{x_i^V\}_{i \in I}$. Indeed, the sets of arguments of histograms $X^U = \{x_i^U\}_{i \in I_U}$ and $X^V = \{x_i^V\}_{i \in I_V}$ are combined: $X = X^{(U)} \cup X^{(V)} = \{x_i\}$ and some procedure for filling data gaps is applied. Thus, we assume below that all histograms are of the form $U = (x_i, u_i)_{i \in I} = (u_i)_{i \in I}$.

3 Some Probabilistic Indices of Comparison

Let $U = (x_i, u_i)_{i \in I}$ and $V = (x_j, v_j)_{j \in I}$ be two histograms, $u_i \geq 0$, $v_j \geq 0$ for all $i, j \in I$, I be some index set.

We consider a numerical index $r(U,V)$ of pairwise comparison of histograms U and V in $\mathcal{U}^2$. We will assume that index $r(U,V)$ is coordinated with the condition, that ordering arguments histograms ascend their importance: if $U = (x_i, u_i)$, $V = (x_i, v_i)$ are two histograms for which $u_i = v_i$ for all $i \neq k, l$ and $u_l - v_l = v_k - u_k \geq 0$ then $r(U,V) \geq 0$ for $k > l$ and $r(U,V) \leq 0$ for $k < l$. Hence, in particular it follows that $r(U,U) = 0$.

If the index $r(U,V)$ is given with help of some utility function $F(U)$ as $r(U,V) = F(U) - F(V)$ then $U \succ V \Leftrightarrow r(U,V) \geq r(V,U) \Leftrightarrow \Delta_r(U,V) = r(U,V) - r(V,U) \geq 0$ will be total preorder relation. In general case the sign of differential index of comparison $\Delta_r(U,V) = r(U,V) - r(V,U)$ cannot assign a transitive relation.

We give examples of indices pairwise comparison of histograms – probability distributions. In this case we assume that $U = (x_i, u_i)_{i \in I}$ and $V = (x_j, v_j)_{j \in I}$ are random variables taking values $\{x_i\}_{i \in I}$ with probabilities $\{u_i\}_{i \in I}$ and $(v_j)_{j \in I}$ accordingly.

1. Let $U \succ V$ if $E[U] \geq E[V]$ (comparison of mathematical expectations). In general $U \succ V$ if $E[f(U)] \geq E[f(V)]$, where f is some function (utility function). We normalize this index that it accepts values in the interval [0,1]: $E_0[U] = \frac{1}{\Delta x}\left(E[V] - x_{\min}\right)$, where $\Delta x = x_{\max} - x_{\min}$. Notice that $E_0[U] = E[U_0]$, where $U_0 = (x_i^0, u_i)_{i \in I}$, $x_i^0 = \frac{1}{\Delta x}\left(x_i - x_{\min}\right) \in [0,1]$ for all $i \in I$. The corresponding differential comparison index is denoted by $\Delta_E(U,V) = E_0[U] - E_0[V] = \frac{1}{\Delta x}\left(E[U] - E[V]\right)$.

2. Let $U \succ V$ if $F_U(x) \leq F_V(x)$ for all $x \in \mathbb{R}$ where $F_U(x) = \sum_{i: x_i < x} u_i$ is distribution function of random variable U. The opposite inequality in the comparison is explained by a condition of conformity of comparison with ordering of arguments of comparing histograms by ascending importance. This is a principle of stochastic dominance of the 1st order, which is used, for example, in the risk theory [13].

The corresponding differential comparison index is denoted by $\Delta_F(U,V) = \inf\limits_x \left(F_U(x) - F_V(x)\right)$. Notice that $F_U(x) - F_V(x) = 0$ for all $x \leq x_1$ or $x > x_n$ if $U = (x_i, u_i)_{i=1}^n$ and $V = (x_j, v_j)_{j=1}^n$ are two random variables. We will consider the index $\inf\limits_{x \in (x_1, x_n]} \left(F_U(x) - F_V(x)\right)$ instead of differential comparison index $\Delta_F(U,V) = \inf\limits_x \left(F_U(x) - F_V(x)\right)$ because the conditions of conservation of sign of difference $F_U(x) - F_V(x)$ are interesting for us. We will denote this index by $\Delta_F(U,V)$ too. Notice that index $\Delta_F(U,V)$ is not defined on the entire set $\mathcal{U}^2$.

3. Let $U \succ V$ if $P\{U \geq V\} \geq P\{U \leq V\}$. This approach to comparison is called stochastic precedence (V precedes U) and the some properties of this ordering can be found in [14,15]. If we assume that the random variables $U = (x_i, u_i)_{i \in I}$ and $V = (x_j, v_j)_{j \in I}$ are independent then $P\{U \geq V\} = \sum\limits_{(i,j):\, x_i \geq x_j} u_i v_j$. The corresponding differential comparison index is denoted by $\Delta_P(U,V) = P\{U \geq V\} - P\{U \leq V\}$. Notice that the inequality $\Delta_P(U,V) \geq 0$ does not specify a transitive relation. However, the probability of nontransitive

triples of histograms for uniform generation is very small as shown by numerical simulation.

4 Distortions of Histograms

Suppose that we have two "distorted" histograms $\tilde{U} = (x_i, \tilde{u}_i)_{i\in I}$ and $\tilde{V} = (x_j, \tilde{v}_j)_{j\in I}$ instead compared histograms $U = (x_i, u_i)_{i\in I}$ and $V = (x_j, v_j)_{j\in I}$. There are different reasons for distortions of histograms. It may be intentional manipulation by histogram data. It may be the result of random factors. It may be the result of the processing procedures of histogram (smoothing, reduction to the unimodal form, etc.). Therefore the description of uncertainty of histogram may be different. For example, this uncertainty may have an interval or stochastic or fuzzy character, etc.

We consider the interval distortions of histograms below. Let $U = (x_i, u_i)_{i\in I}$ is a ideal histogram and $\tilde{U} = (x_i, \tilde{u}_i)_{i\in I}$ is an interval distortion of U: $\tilde{u}_i = u_i + h_i$, $i \in I$, where $\sum_{i\in I} h_i = 0$ and $|h_i| \leq \alpha u_i$, $i \in I$, where $\alpha \in [0, 1]$. The value α characterize the threshold of distortion. We will call such a distortion an α-distortion. We denote by $N_\alpha(U)$ the class of all α-distortion of histogram $U = (x_i, u_i)_{i\in I}$, i.e.

$$N_\alpha(U) = \left\{H = (h_i)_{i\in I} : \sum\nolimits_{i\in I} h_i = 0,\ |h_i| \leq \alpha u_i,\ i \in I\right\}. \tag{1}$$

Suppose that $\Delta_r(U, V) \geq 0$. The main question that is studied in this paper consist in following. In what case do we have $\Delta_r(\tilde{U}, \tilde{V}) \geq 0$ for all $H \in N_\alpha(U)$ and $G \in N_\beta(V)$? By other words, when the comparison of histograms will not change after α-distortion of histogram $U = (x_i, u_i)_{i\in I}$ and β-distortion of histogram $V = (x_j, v_j)_{j\in I}$? We obtain the conditions of conservation of comparison distorted histograms for different types of comparisons.

5 Conditions of Preservation for Comparison of Distorted Histograms

The Conservation Conditions of Comparison of Distorted Histograms with Respect to $\boldsymbol{\Delta_E}$ Index. We consider the value

$$\mathcal{E}_U = \sup\left\{\sum\nolimits_{i\in I} x_i^0 h_i : (h_i)_{i\in I} \in N_1(U)\right\}$$

for histogram $U = (x_i, u_i)_{i\in I}$, where $N_1(U)$ is a set of the type (1) with $\alpha = 1$. We note the following properties of the value $\mathcal{E}_U$.

Lemma 1. *The estimation* $0 \leq \mathcal{E}_U \leq \min\{E_0[U], 0.5\}$ *is true and this inequality is sharp.*

Lemma 2. *The equality*

$$\mathcal{E}_U = \sum_{s=s_0}^{n} x_s^0 u_s a_s - \sum_{s=1}^{s_0-1} x_s^0 u_s b_s$$

is true for histogram $U = (x_i, u_i)_{i=1}^n$*, where* $1 \geq a_n \geq ... \geq a_{s_0} \geq 0$*,* $1 \geq b_1 \geq ... \geq b_{s_0-1} \geq 0$*,* $\sum_{s=s_0}^n u_s a_s = \sum_{s=1}^{s_0-1} u_s b_s$*, and the index* s_0 *satisfies to inequality* $s_0 - 1 < m_U \leq s_0$*, where* m_U *is a median of distribution of* U*.*

Proposition 3. *Let* $\tilde{U} = (x_i, u_i + h_i)_{i \in I}$*,* $\tilde{V} = (x_j, v_j + g_j)_{i \in I}$ *be a* α*- and* β*-distortion of histograms* $U = (x_i, u_i)_{i=1}^n$ *and* $V = (x_j, v_j)_{j=1}^n$ *respectively. Then we have* $\Delta_E(\tilde{U}, \tilde{V}) \geq 0$ *for all* $(h_i)_{i \in I} \in N_\alpha(U)$ *and* $(g_i)_{i \in I} \in N_\beta(V)$*,* $\alpha, \beta \in [0, 1]$ *iff* $\Delta_E(U, V) \geq \alpha \mathcal{E}_U + \beta \mathcal{E}_V$*.*

Let $\bar{\mathcal{E}}_U = \min\{E_0[U], 0.5\}$. Then following corollary follows from Lemma 1.

Corollary 4. *If we have* $\Delta_E(U, V) \geq \alpha \bar{\mathcal{E}}_U + \beta \bar{\mathcal{E}}_V$*, then inequality* $\Delta_E(\tilde{U}, \tilde{V}) \geq 0$ *is true for all* $(h_i)_{i \in I} \in N_\alpha(U)$ *and* $(g_i)_{i \in I} \in N_\beta(V)$*.*

The Conservation Conditions of Comparison of Distorted Histograms with Respect to Δ_F Index. The similar conditions of the conservation of a sign of the comparison can be obtained for differential index $\Delta_F(U, V)$. We introduce the function

$$\mathcal{F}_U(x) = \sup\left\{\sum\nolimits_{i: x_i < x} h_i : \ (h_i)_{i \in I} \in N_1(U)\right\},$$

where $N_1(U)$ is a set of type (1) with $\alpha = 1$.

Lemma 5. $\mathcal{F}_U(x) = \min\{F_U(x), 1 - F_U(x)\}$ *for all* $x \in \mathbb{R}$*.*

Proposition 6. *Let* $\tilde{U} = (x_i, u_i + h_i)_{i \in I}$*,* $\tilde{V} = (x_j, v_j + g_j)_{i \in I}$ *be a* α*- and* β*-distortion of histograms* $U = (x_i, u_i)_{i \in I}$ *and* $V = (x_j, v_j)_{i \in I}$ *respectively. Then we have* $\Delta_F(\tilde{U}, \tilde{V}) \geq 0$ *for all* $(h_i)_{i \in I} \in N_\alpha(U)$ *and* $(g_i)_{i \in I} \in N_\beta(V)$*,* $\alpha, \beta \in [0, 1]$ *if*

$$F_U(x) - F_V(x) \geq \alpha \mathcal{F}_U(x) + \beta \mathcal{F}_V(x) \ \text{ for all } \ x \in \mathbb{R}.$$

Corollary 7. *The inequality* $\Delta_F(\tilde{U}, \tilde{V}) \geq 0$ *is true for all* $(h_i)_{i \in I} \in N_\alpha(U)$ *and* $(g_i)_{i \in I} \in N_\beta(V)$ *if* $0 \leq \sup\limits_x \frac{\alpha \mathcal{F}_U(x) + \beta \mathcal{F}_V(x)}{F_U(x) - F_V(x)} \leq 1$ *(we assume that the fraction is equal to zero if its numerator and denominator are equal to zero).*

Corollary 8. *If* $\Delta_F(U, V) \geq \sup\limits_x \{\alpha \mathcal{F}_U(x) + \beta \mathcal{F}_V(x)\}$ *then inequality* $\Delta_F(\tilde{U}, \tilde{V})$ ≥ 0 *is true for all* $(h_i)_{i \in I} \in N_\alpha(U)$ *and* $(g_i)_{i \in I} \in N_\beta(V)$*.*

The Conservation Conditions of Comparison of Distorted Histograms with Respect to Δ_P Index. The following conditions of sign conservation are valid for differential comparison index $\Delta_P(U, V)$.

Proposition 9. *Let $\tilde{U} = (x_i, u_i + h_i)_{i \in I}$, $\tilde{V} = (x_j, v_j + g_j)_{j \in I}$ be a α- and β-distortion of histograms $U = (x_i, u_i)_{i \in I}$ and $V = (x_j, v_j)_{j \in I}$ respectively. Then we have $\Delta_P(\tilde{U}, \tilde{V}) \geq 0$ for all $(h_i)_{i \in I} \in N_\alpha(U)$ and $(g_i)_{i \in I} \in N_\beta(V)$, $\alpha, \beta \in [0, 1]$ if $\Delta_P(U, V) \geq \Delta\eta_{\alpha,\beta}(U, V)$, where*

$$\Delta\eta_{\alpha,\beta}(U, V) = \sup_{\substack{(h_i)_i \in N_\alpha(U), \\ (g_i)_i \in N_\beta(V)}} \sum_{(i,j):\, x_i < x_j} (u_i g_j + h_i v_j + h_i g_j - u_j g_i - h_j v_i - h_j g_i).$$

6 Comparison of the Sets of Admissible Distortions

We consider the set of all those α- and β-distortion of histograms U and V respectively that preserve the histogram comparison with respect to given index $\Delta_r(U, V)$ on condition that it equals $c > 0$:

$$\Omega_r^c(U, V) = \left\{ (\alpha, \beta) :\ \Delta_r(U, V) = c,\ \Delta_r(\tilde{U}, \tilde{V}) \geq 0\ \forall H \in N_\alpha(U),\ G \in N_\beta(V) \right\}.$$

This set is called the set of admissible distortions of histograms U and V for given comparison $\Delta_r(U, V) = c$. It is easy to see that the set $\Omega_r^c(U, V)$ is a star domain (or star-convex set, star-shaped or radially convex set) [16] with star center the origin, i.e. if $(\alpha_0, \beta_0) \in \Omega_r^c(U, V)$ then $(t\alpha_0, t\beta_0) \in \Omega_r^c(U, V)$ for all $t \in [0, 1]$. It is known [16] that ray function $\Phi_r^c(\alpha, \beta)$ (i.e. continuous, non-negative and homogeneous: $\Phi_r^c(t\alpha, t\beta) = t\Phi_r^c(\alpha, \beta)$ for all $t \geq 0$) may be set in bijective correspondence to star-convex set with center at the origin such that $\Omega_r^c(U, V) = \{(\alpha, \beta) :\ \alpha \geq 0, \beta \geq 0, \Phi_r^c(\alpha, \beta) \leq 1\}$.

The functions $\Phi_E^c(\alpha, \beta)$, $\Phi_F^c(\alpha, \beta)$ and $\Phi_P^c(\alpha, \beta)$ of sets of admissible distortions for indices $\Delta_E(U, V)$, $\Delta_F(U, V)$ and $\Delta_P(U, V)$ respectively will be equal $\Phi_E^c(\alpha, \beta) = \frac{1}{c}\left(\alpha\mathcal{E}_U + \beta\mathcal{E}_V\right)$, $\Phi_F^c(\alpha, \beta) = \sup\limits_x \left\{ \frac{\alpha\mathcal{F}_U(x) + \beta\mathcal{F}_V(x)}{F_U(x) - F_V(x)} \right\}$, $\Phi_P^c(\alpha, \beta) = \frac{1}{c}\Delta\eta_{\alpha,\beta}(U, V)$ as follows from the Proposition 3, 6, 9.

In general the function $\Phi_F^c(\alpha, \beta)$ is a piecewise linear in the case of discrete distributions. However we can specify the wide class of pairs of distributions for which this function is more simple. Let U and V be two random variables with distribution functions F_U and F_V respectively, m_U and m_V be a medians of corresponding distributions. If the values $F_U(m_V)$ and $F_V(m_U)$ are approximately symmetrical with respect to $\frac{1}{2}$ then the function $\Phi_F^c(\alpha, \beta)$ consists of two linear functions. The function $\Phi_F^c(\alpha, \beta)$ is linear function if the values $F_U(m_V)$ and $F_V(m_U)$ are located "strongly asymmetric" with respect to $\frac{1}{2}$.

We introduce the following notion for numerical measuring the degree of stability of the comparison to the α-distortion. We call the comparison $r(U, V)$ of histograms U and V with $r(U, V) = c > 0$ by δ-stable to α-distortion if $\delta = \max\{k(\alpha, \beta) :\ \Phi_r^c(\alpha, \beta) \leq 1\}$, where $k(\alpha, \beta)$ is a some criterial function, as which the may be, for example: $k_1(\alpha, \beta) = \frac{1}{2}(\alpha + \beta)$, $k_2(\alpha, \beta) = \min\{\alpha, \beta\}$.

By other words, δ-stability characterizes the maximal level of distortions of histograms for which the sign of comparison histograms will not change. We denote the value of δ-stability of comparison of histograms $r(U,V)$ relatively criterial function k_i through $\delta_r^{(i)}(U,V)$. In particular, it is easy to see that $\delta_E^{(1)}(U,V) = \frac{c}{2\min\{\mathcal{E}_U,\mathcal{E}_V\}}$, $\delta_E^{(2)}(U,V) = \frac{c}{\mathcal{E}_U+\mathcal{E}_V}$.

Example. We consider the comparison of the two histograms of USE (Unified State Exam) applicants admitted in 2012 on a specialty "Economy" and only on the competitive set in Moscow State Institute of the International Relations (MGIMO, the histogram U) and Moscow State University (MSU, the histogram V). The histograms of these universities are given in Fig. 1.

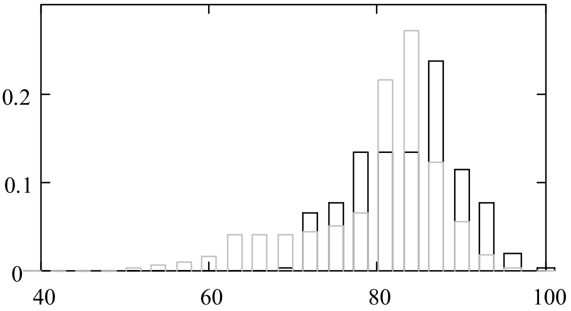


Fig. 1. The histograms USE applicants admitted in 2012 on a specialty "Economy" in Moscow State Institute of the International Relations (dark color) and Moscow State University (light color)

The normalized expectations have values $E_0[U] = 0.732$ and $E_0[V] = 0.669$ for these histograms; the differential index of comparison with respect to expectations is equal $\Delta_E(U,V) = E_0[U] - E_0[V] = 0.063$; the differential index of comparisons with respect to distribution functions is equal $\Delta_F(V,U) = \inf\limits_{x\in(x_1,x_n]} (F_V(x) - F_U(x)) = 0.0031$; we have probabilities $P\{U \geq V\} = 0.684$, $P\{U \leq V\} = 0.434$ and the differential index of comparisons with respect to probabilities is equal $\Delta_P(U,V) = P\{U \geq V\} - P\{U \leq V\} = 0.25$.

Then we have following values of δ-stability of comparisons of histograms with respect to:

a) expectations: $\delta_E^{(1)}(U,V) = 0.375$, $\delta_E^{(2)}(U,V) = 0.351$;

b) distribution functions: $\delta_F^{(1)}(U,V) = 0.001989$; $\delta_F^{(2)}(U,V) = 0.001788$;

c) probabilities: $\delta_P^{(1)}(U,V) = 0.306$, $\delta_P^{(2)}(U,V) = 0.254$.

Thus the comparisons with respect to expectation shows the greatest stability (at the level of 35-40%). The comparisons with probability slightly worse than the first comparison (25-30%). The comparison using the distribution function has the lowest stability (0.15-0.20%).

7 Conclusion

The necessary and sufficient conditions on the distortion level of histograms, under which the result of the comparison of histograms by probabilistic methods will not change, were found in this paper. It was clear a priori that ”integral” methods of comparison, such as the method of comparing expectations, method comparisons of probability of inequalities are more preferred than pointwise comparison methods, such as stochastic dominance. These assumptions were confirmed by the results of research. Accurate theoretical estimates of possible values of distortion histograms, in which the comparison result will not change, were obtained.

The found conditions invariability of comparing histograms can be used to estimate the reliability of results of different rankings, data processing, etc., in terms of different types of uncertainty: stochastic uncertainty, the uncertainty associated with the distortion of the data in filling data gaps, etc.

References

1. Shnoll, S.E., Zenchenko, K.I., Udaltsova, N.V.: Cosmophysical Effects in the Structure of Daily and Yearly Periods of Changes in the Shape of Histograms Constructed from the Measurements of 239P u alpha-Activity. Biophysics 49(1), 155–155 (2004)
2. Aleskerov, F.T., Belousova, V., Serdyuk, M., Solodkov, V.M.: Dynamic Analysis of the Behavioural Patterns of the Largest Commercial Banks in the Russian Federation. Working papers by International Centre for Economic Research. Series “Applied Mathematics Working Paper Series” 12 (2008)
3. Vanegas, L.V., Labib, A.W.: Application of New Fuzzy-Weighted Average (NFWA) Method to Engineering Design Evaluation. Int. J. Prod. Res. 39, 1147–1162 (2001)
4. Fodor, J., Roubens, M.: Fuzzy Preference Modelling and Multicriteria Decision Support. Kluwer Academic Publishers, Dordrecht (1994)
5. Shorrocks, A.F.: Ranking Income Distributions. Economica 50, 3–17 (1983)
6. Bobrov, R.A., Lepskiy, A.E.: Ranking Universities According to the Results of USE by Means of Fuzzy Numbers Comparison Methods. Working paper WP7/2014/01, Publishing House of the Higher School of Economics, Moscow (2014) (in Russian)
7. Podinovski, V.V.: Criteria Importance Theory. Math. Soc. Sci. 27, 237–252 (1994)
8. Aleskerov, F.T., Chistyakov, V.V., Kaliaguine, V.A.: Social Threshold Aggregations. Social Choice and Welfare 35(4), 627–646 (2010)
9. Wang, X., Ruan, D., Kerre, E.E.: Mathematics of Fuzziness - Basic Issues. Springer, Heidelberg (2009)
10. Baas, S.M., Kwakernaak, H.: Rating and Ranking of Multiple-Aspect Alternatives Using Fuzzy Sets. Automatic 13, 47–58 (1977)
11. Yager, R.R.: A Procedure for Ordering Fuzzy Sets of the Unit Interval. Inf. Sciences 24, 143–161 (1981)
12. Dubois, D., Prade, H.: Ranking Fuzzy Numbers in the Setting of Possibility Theory. Inf. Science 30, 183–224 (1983)
13. Wolfstetter, E.: Topics in Microeconomics: Industrial Organization, Auctions, and Incentives. Cambridge Univ. Press, Cambridge (1999)
14. Boland, P.J., Singh, H., Cukic, B.: The Stochastic Precedence Ordering with Applications in Sampling and Testing. J. of Applied Probability 41(1), 73–82 (2004)
15. De Santis, E., Fantozzi, F., Spizzichino, F.: Relations between Stochastic Orderings and Generalized Stochastic Precedence (2014), `http://arxiv.org/pdf/1307.7546.pdf`
16. Cassels, J.W.S.: An Introduction to the Geometry of Numbers. Springer, Heidelberg (1959)

The Fuzzy Representation of Prior Information for Separating Outliers in Statistical Experiments

Dmitry A. Matsypaev[1] and Andrey G. Bronevich[2]*

[1] Southern Federal University, Taganrog, Russia
[2] National Research University Higher School of Economics, Moscow, Russia
{dmitry.matsypaev,andreybronevich}@gmail.com

Abstract. The paper presents a new fuzzy set based description which helps to distinguish the expected values of the statistical experiment from the outliers. Since the Neyman-Pearson criterion is not adequate in some real applications for such purpose, we propose to use triangular norms for conjuction of two propositions about typical and non-typical values and describe both of them as a fuzzy set that is called the typical transform. We also investigate such a property of the typical transform as stability.

Keywords: distortion function, triangular norm, fuzzy set, Neyman-Pearson criterion, outliers, Lipschitz continuity.

1 Introduction

In general when we carry out statistical experiments we can observe results that are not expected. These results are called outliers. To the contrary results that are conceived expected can be called typical. The paper gives a new fuzzy set based description of this prior information that we call the typical transform. For this purpose we consider two probability measures. The first measure describes the typical elementary events that can appear during the experiment and the second measure describes the whole possible events during the experiment, i.e. the probability distribution that can be chosen if we know nothing about the possible outcomes of the experiment. This probability measure can be chosen using, for example, the maximal entropy principle that leads to the uniform distribution.

In the paper we argue that the Neyman-Pearson criterion is not adequate in some real applications for separating typical and non-typical results of the experiment, therefore, we propose to use t-norms for conjunction of two propositions about typical values that found applications in fuzzy set theory. This allows us to describe typical and non-typical results of the experiment with a fuzzy set that is called the typical transform. We investigate some properties of the typical transform, in particular, its stability.

* This study (research grant No 14-01-0015) was supported by The National Research UniversityHigher School of Economics Academic Fund Program in 2014/2015.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*, Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_20

2 Background

The detection of outliers is purely heuristic-based problem because there are no exact definition of that term. That is why there are a lot of approaches which are based on various assumptions[4]. The methods based on statistical tests assume that the data is distributed normally and identify outliers as the observations which are far enough from the mean in terms of standard deviation or Mahalanobis distance. Depth-based approaches assume that outliers are located at the border of the data space. Deviation-based methods state that outliers are the observations whose removal minimizes the variance of the data. Distance-based approaches assume that outliers are far apart from their nearest neighbours in the data set. Density-based methods estimate the density of probability distribuition for certain observation point and compare it with densities of the nearest neighboring observations.

Our approach based on the typical transform assumes that the probability density function is specified. We formulate three general postulates about the relation between the probability density function value and the corresponding degree of typicality. Non-typical elementary events are treated as outliers.

The typical transform can also be considered as a new fuzzy methodology to describe the uncertainty during the processing of the statistical information. One should note the existence of other approaches to build the fuzzy sets when one processes the statistical information[1,3].

3 Necessary Mathematical Apparatus

In the next section we use *distortion functions* [2,7] and *triangular norms* [5,6] to proceed our reasoning.

Let $\mathscr{F}$ be the set of all distortion functions. According to the definition, the arbitrary distortion function $\phi : [0,1] \to]0,1]$ is a non-decreasing function that satisfies $\phi(0) = 0$ and $\phi(1) = 1$.

Let $\mathscr{T}$ be the set of all triangular norms. Each triangular norm $t : [0,1] \times [0,1] \to [0,1]$ satisfies the following conditions:

- commutativity $t(a,b) = t(b,a)$;
- monotonicity $t(a,b) \geq t(c,d)$ if $a \geq c$ and $b \geq d$;
- associativity $t(a,t(b,c)) = t(t(a,b),c)$;
- identity element $t(a,1) = a$.

4 Typical Transform

4.1 Problem Statement

Let (U, σ_U, P_U) be a probability space, where $U \subseteq \mathbb{R}^n$ is the set of elementary events, σ_U is the sigma algebra of measurable subsets in U, and P_U is a probability measure on σ_U. Let us assume that the measure P_U plays a role of vacuous information about the experiment and can be chosen, for example, by using the maximum entropy principle. In case of additional information about the experiment we describe it with a random variable $\rho : U \to \mathbb{R}$ and corresponding probability measure P_ρ with a density function $p_\rho : U \to [0, \infty)$. Thus the probability of any $A \in \sigma_U$ can be computed by the formula $P_\rho(A) = \int_A p_\rho(u)du$, where the last integral can be conceived as the Lebesgue integral in general case. For any random variable ρ we divide its values on typical and non-typical ones using the set of typical events $B_\rho \in \sigma_U$. This separation heuristically can be described by the following postulates.

Postulate 1. *Let $p_\rho(u_1) > p_\rho(u_2)$ for $u_1, u_2 \in U$, then u_1 is more likely typical than u_2.*

Since the proceeded experiment changes the probability distribution over the U, the regions with a high degree of outcomes condensation tend to contain a lot of posterior probability and a little of prior probability.

Postulate 2. *The value $P_U(B_\rho)$ should be close to* 0.

Postulate 3. *The value $P_\rho(B_\rho)$ should be close to* 1.

4.2 The Formalization of Postulates

Consider an arbitrary elementary event $u \in U$. Let us introduce a pair of hypotheses which are collectively exhaustive events. The main hypothesis H_0 states that the elementary event u is typical for the random variable ρ

$$H_0 : u \in B_\rho.$$

The alternative hypothesis H_1 says the opposite, i.e. u is not typical for the random variable ρ

$$H_1 : u \notin B_\rho.$$

Let us define a random variable χ such that its probability density function depends on the choice between the hypotheses H_0 and H_1, i.e. $p_\chi(u; H_0) \neq p_\chi(u; H_1)$.

Due to the postulate 1, the optimal statistical criterion to classify elementary events should have the critical region like

$$S_{\rho,y} = \{u \in U | p_\rho(u) < y\}, \tag{1}$$

where $y \geq 0$ is the parameter.

Consider the probability of type I error $\alpha(y)$ and the probability of type II error $\beta(y)$ for the parametrized criterion with the critical region $S_{\rho,y}$:

$$\alpha(y; \rho, \chi) = \int_{S_{\rho,y}} p_\chi(u; H_0)du,$$

$$\beta(y; \rho, \chi) = \int_{U \setminus S_{\rho,y}} p_\chi(u; H_1)du.$$

The postulate 2 can be formalized as the minimization of the following functional

$$\int_{U \setminus S_{\rho,y}} dP_U \to \min_y . \tag{2}$$

The postulate 3 can be formalized as the minimization of the functional

$$\int_{S_{\rho,y}} dP_\rho \to \min_y . \tag{3}$$

Note that if $p_\chi(u; H_0) = p_\rho(u)$ then (3) minimizes the probability of type I error $\alpha(y)$. Analogously, if $p_\chi(u; H_1)du = dP_U$, then (2) minimizes the probability of type II error $\beta(y)$.

Thus, the main H_0 and alternative H_1 hypotheses are presented. The critical region for the optimal statistical criterion to check the main hypothesis belongs to the parametrized family (1). The optimal criterion should minimize probabilitiies of type I and type II errors that can be presented as the system of conditions

$$\begin{cases} \alpha(y;\rho) = \displaystyle\int_{S_{\rho,y}} dP_\rho \to \min_y , \\ \beta(y;\rho) = \displaystyle\int_{U \setminus S_{\rho,y}} dP_U \to \min_y . \end{cases} \tag{4}$$

The system (4) is the multi-objective optimization problem. According to (4), type I $\alpha(y)$ and type II $\beta(y)$ errors can not be optimized simultaneously. The Neyman-Pearson criterion suggests to fix an admissible level of the type I error probability and build the optimal criterion that minimizes the type II error probability given the fixed type I error level. In the context of the problem being discussed, it follows the priority of the postulate 3 over the postulate 2. Assuming the equivalence of the postulates 3 and 2, we propose another way to solve the multi-objective problem (4) which takes into account the specificity of present restrictions.

Assume that the statistical criterion with the critical region $S_{\rho,y}$ from the criteria family (1) satisfies the postulate 3 with the confidence degree

$$C_\alpha(y;\rho,\phi_\alpha) = \phi_\alpha(1 - \alpha(y;\rho)) = \phi_\alpha\left(\int_{U \setminus S_{\rho,y}} dP_\rho\right),$$

where $y \geq 0$ and $\phi_\alpha \in \mathscr{F}$. Analogously, let the same criterion satisfy the postulate 2 with the confidence degree

$$C_\beta(y;\rho,\phi_\beta) = \phi_\beta(1 - \beta(y;\rho)) = \phi_\beta\left(\int_{S_{\rho,y}} dP_U\right),$$

where $y \geq 0$ and $\phi_\beta \in \mathscr{F}$. In both cases the confidence degree depends on the error of certain type transformed by the distortion function which is known a priori. The logical conjunction of the postulates 2 and 3 can be expressed with the help of the triangular norm $t \in \mathscr{T}$

$$C(y;\rho,\phi_\alpha,\phi_\beta,t) = t\left(\phi_\alpha\left(\int_{U\setminus S_{\rho,y}} dP_\rho\right), \phi_\beta\left(\int_{S_{\rho,y}} dP_U\right)\right), \tag{5}$$

which is known a priori too.

The maximal confidence degree of both postulates is achieved when $y = y^*$, where

$$y^* = \arg\max_{y\geq 0} C(y;\rho,\phi_\alpha,\phi_\beta,t), \tag{6}$$

which corresponds to the critical region

$$S_{\rho,y^*}(u) = \{u \in U | p_\rho(u) < y^*\}. \tag{7}$$

If $u \in U$ belongs to the critical region S_{ρ,y^*} then the alternative hypothesis H_1 is accepted. Otherwise, the main hypothesis H_0 is admitted. Thus, the elementary event $u \in U$ is typical $u \in B_\rho$ if it does not belong to the critical region S_{ρ,y^*}. This statement can be formalized in the terms of characteristic functions of the sets B_ρ and S_{ρ,y^*}

$$B_\rho(u;\phi_\alpha,\phi_\beta,t) = 1 - S_{p,y^*}(u). \tag{8}$$

4.3 Fuzzy Set of Typical Elementary Events

The solution (6)-(8) can be unstable relative to the small changes in the function p_ρ since the output of characteristic function is binary valued. However, (6)-(8) is naturally extended to the case if the critical region is a fuzzy set. Assuming that S is the fuzzy critical region, the main and alternative hypotheses for $u \in U$ are accepted with $1 - S(u)$ and $S(u)$ confidence degrees respectively. Thus, the set of the typical elementary events B_ρ in (8) can be treated as a fuzzy set.

If $B_\rho(u)$ is close to 1 or 0 then u is accordingly almost certainly the typical or almost certainly the non-typical elementary event. Meanwhile, if $B_\rho(u)$ is close to 0.5, then there is not enough information to classify u with the high degree of assurance. At this point, the fuzzy sets of typical elementary events are applicable in the case one needs to filter out the elementary events that can not be classified with required degree of confidence.

Let us build the fuzzy critical region as the weighted union of the crisp critical regions $S_{\rho,y}$ using the functional (5)

$$S(u;\rho,\phi_\alpha,\phi_\beta,t,h) = \frac{\int_0^\infty C(y;\rho,\phi_\alpha,\phi_\beta,t)S_{\rho,y}(u)h(y)dy}{\int_0^\infty C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy},$$

where $y \geq 0$ and $h(y) : [0,\infty) \to [0,\infty)$ is the a priori known probability density function which guarantees the convergence of the integrals both in numerator and deniminator of the functional above. Thus, the membership function of the typical elementary events set B_ρ for the random variable ρ can be represented as

$$B_\rho(u;\phi_\alpha,\phi_\beta,t,h)=\frac{\int_0^{p_\rho(u)}C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy}{\int_0^\infty C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy}. \tag{9}$$

4.4 Stability Research

The stability relative to the small changes in the input data is the significant feature in practice. Let us formulate the stability conditions of the transform (1),(5),(9) as the theorem below.

Theorem 1. *Let $\phi_\alpha,\phi_\beta\in\mathscr{F}$ and $t\in\mathscr{T}$ satisfy the Lipschitz [5,6] continuity condition:*

$$|\phi_\alpha(x_1)-\phi_\alpha(x_2)|\le K_\alpha|x_1-x_2|,|\phi_\beta(x_1)-\phi_\beta(x_2)|\le K_\beta|x_1-x_2|,$$
$$|t(x_1,y_1)-t(x_2,y_2)|\le K_t(|x_1-x_2|+|y_1-y_2|)$$

for all $x_1,x_2,y_1,y_2\in[0,1]$. Assume the function h satisfy the conditions

$$\int_0^\infty\frac{h(y)}{y}dy=K_h<\infty,\ \max_{y\in[0;\infty)}h(y)\le h_{max}<\infty.$$

Let ρ,κ be the random variables and $|p_\rho(u)-p_\kappa(u)|\le\varepsilon_{max}$ for all $u\in U$, where $\varepsilon_{max}\ge0$. Then

$$|B_\rho(u;\phi_\alpha,\phi_\beta,t,h)-B_\kappa(u;\phi_\alpha,\phi_\beta,t,h)|\le\frac{K_{max}\varepsilon_{max}}{\int_0^\infty C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy},$$

where $K_{max}=2K_t(K_\beta h_{max}+K_\alpha h_{max}+K_\alpha K_h)+h_{max}$.

Theorem 1 states that the stability of the transform(1),(5),(9) is in the direct proportion with the value

$$q_\rho=\int_0^\infty C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy.$$

If $q_\rho\to0$, then the small changes in the input data may lead to the huge changes of the values of the membership function B_ρ. In the extreme case $q_\rho=0$ the result of (9) is undefined. For instance, that is achieved when the random variable ρ is distributed uniformly on the bounded space U.

We propose to regularize (9) as follows:

$$B_\rho(u;\phi_\alpha,\phi_\beta,t,h)=0.5(1-\lambda)+\lambda\frac{\int_0^{p_\rho(u)}C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy}{\int_0^\infty C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy}. \tag{10}$$

The value $\lambda\in[0,1]$ is the regularization parameter which is chosen according to the stability of (9), i.e. it is directly proportional to q_ρ:

$$\lambda=\min\left\{1,\frac{E_{max}\int_0^\infty C(y;\rho,\phi_\alpha,\phi_\beta,t)h(y)dy}{K_{max}}\right\}, \tag{11}$$

where E_{max} is the number that bounds from above the absolute value of $\varDelta B_\rho$ for small changes in the input data

$$\frac{|\varDelta B_\rho(u)|}{\varepsilon_{max}} \leq E_{max}.$$

One can prove that if the functions ϕ_α, ϕ_β, t and h satisfy the conditions of the theorem 1, then (10) is absolutely stable regardless of q_ρ for all $u \in U$.

The unstability of (9) when $q_\rho \to 0$ can be treated as the high degree of uncertainty whether $u \in U$ is typical elementary event ot not. One states above that if closer the membership degree of B_ρ to 0.5 for some $u \in U$ then greater the uncertainty degree during its classification as typical or not. The formulas (10) and (11) naturally connect together both these uncertainty types. The closer q_ρ is to 0, the less the parameter λ is and the closer the membership degree of B_ρ is to 0.5 for all $u \in U$.

4.5 Practical Use

The practical use of the presented transform is to define the degree of how typical the certain signal relative to the statistical experiment proceeded earlier. Let the signal be represented as the probability density function $f : U \to [0, \infty)$ that corresponds to the probability measure P_f. Then the typicality degree for f relative to ρ can be represented as

$$B_\rho(f) = \int_U B_\rho(u) f(u) du = \int_U B_\rho(u) dP_f.$$

The significant feature is the ability to estimate the confidence degree of the introduced transform:

$$T_\rho = \int_U |1 - 2B_\rho(u)| dP_U.$$

One can also compute the confidence degree of this transform applying to the specified signal

$$T_\rho(f) = \int_U |1 - 2B_\rho(u)| dP_f.$$

5 Examples

Here are some examples of the discrete transform when $U = \{-100, \ldots, 100\}$ and P_U has the uniform distribution on U. We use $\phi_\alpha(u) = \phi_\beta(u) = u$ for all $u \in \{-100, \ldots, 100\}$ and $t(x, y) = \min(x, y)$. One can easily prove that they satisfy the Lipschitz condition. The various values of E_{max} were chosen: 1, 15 and 50. See the Fig. 1 for the results of transform.

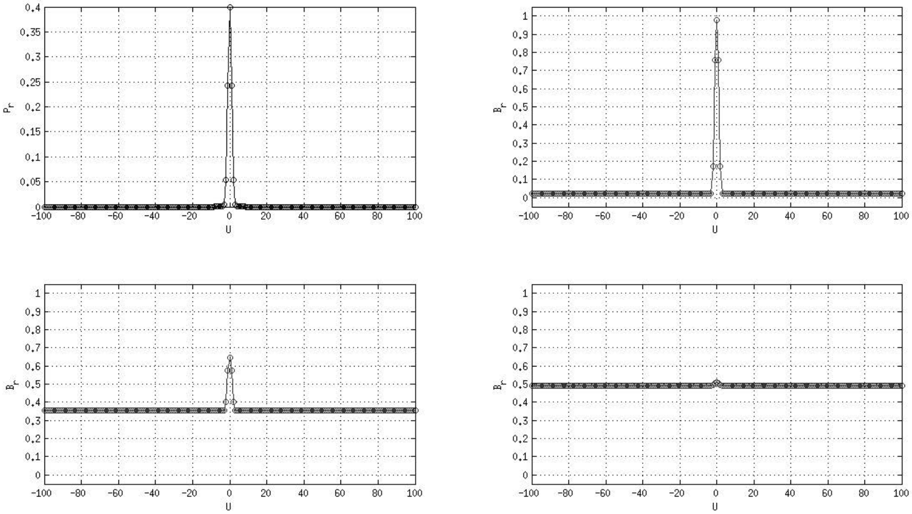


Fig. 1. Top-left item is the original distribution P_{ρ}. Top-right, down-left and down-right items are the results of discrete transform B_{ρ} with $E_{max} = 50$, $E_{max} = 15$ and $E_{max} = 1$ respectively.

6 Conclusion

We present the fuzzy set based description for the statistical experiment to distinguish the expected values from the outliers. We propose to use triangular norms for conjuction of two propositions about typical and non-typical values and describe both of them as the fuzzy set that is called the typical transform. We also investigate such a property of the typical transform as stability.

References

1. Bronevich, A.G., Karkishchenko, A.N.: Statistical classes and fuzzy set theoretical classification of probability distributions. In: Statistical Modeling, Analysis and Management of Fuzzy Data, pp. 173–198. Physica-Verl., Heidelberg (2002)
2. Chateauneuf, A.: Decomposable capacities, distorted probabilities and concave capacities. Mathematical Social Sciences 31, 19–37 (1996)
3. Dubois, D., Prade, H.: Fuzzy sets and probability: misunderstandings, bridges and gaps In. In: Proc. of the Second IEEE Conderence on Fuzzy Systems, pp. 1059–1068. IEEE (1993)
4. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22(2), 85–126 (2004)
5. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms, 1st edn., p. 387. Springer, Heidelberg (2000)
6. Mesiarova, A.: Triangular norms and k-Lipschitz property. In: EUSFLAT Conf., pp. 922–926 (2005)
7. Wallner, A.: Bi-elastic neighbourhood models. In: Proc. of the 3rd International Symposium on Imprecise Probabilities and Their Applications, Lugano, Switzerland, pp. 591–605 (2003)

Similarity Test for the Expectation of a Random Interval and a Fixed Interval

Ana Belén Ramos-Guajardo

Department of Statistics, Operational Research and Mathematic Didactics,
University of Oviedo, Spain
ramosana@uniovi.es

Abstract. A hypothesis test for analyzing the degree of similarity between the expected value of a random interval and a fixed interval is introduced. It is based on a measure of the similarity between classical convex sets proposed in the literature. Asymptotic techniques are firstly applied to analyze the limit distribution of the proposed test statistic. Afterwards, a bootstrap approach is presented to better approximate the sampling distribution. Finally, the performance of the test is investigated by means of simulation studies.

Keywords: Similarity degree, expected value, random interval, bootstrap approach.

1 Introduction

Different problems involving interval-valued data have been faced in the literature. Sometimes intervals are referred to an imprecise identification of an exact value quantification [7,13]. In other situations the interest is focused on characteristics which are essentially interval-valued data as, for instance, fluctuations in Economy, numerical ranges, subjective perceptions and so on [2,3,6].

Random experiments involving such kind of data, also called random intervals (RIs), are considered. Some statistical analysis for RIs in different settings have already been addressed [5,11,12,15]. Specifically, hypothesis tests for the expected value of random intervals, which is also an interval, have been previously developed [8,10]. As these hypotheses comprised strict equalities, the idea is to relax them, which is in coherence with the imprecise setting of intervals.

The aim is to analyze if the expected value of a random interval can be considered to be *similar* to a previously fixed interval. For this purpose, a similarity measure between two intervals which is based on the Jaccard similarity coefficient for classical convex sets [9] will be considered. This index is a ratio quantifying the size of the intersection with respect to the size of the union of both intervals. In this context, the size of the intersection between intervals will be defined taking into account its Lebesgue measure [14], whereas the size of the union is based on the sizes of both intervals and the intersection interval.

A test statistic will be defined on the basis of the Jaccard index and its asymptotic limit distribution will be firstly analyzed. Later, bootstrap techniques

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_21

will be applied in order to approximate the sampling distribution in practice. Some simulation studies will be carried out to show the empirical behaviour of the bootstrap approach.

2 Preliminaries

Let $\mathcal{K}_c(\mathbb{R})$ denote the family of all non-empty closed and bounded intervals of $\mathbb{R}$. The formalization of the hypothesis test procedure is based on the $(\mathrm{mid}\,, \mathrm{spr}\,)$ representation of the intervals, i.e. $A = [\mathrm{mid}\, A \pm \mathrm{spr}\, A]$ for $A \in \mathcal{K}_c(\mathbb{R})$, where $\mathrm{mid}\, A \in \mathbb{R}$ is the mid-point or centre and $\mathrm{spr}\, A \geq 0$ is the spread or radius of A. The previous characterization has been shown to be a valuable tool for different statistical purposes (see, for instance, [2,4,15]).

Given a probability space $(\Omega, \mathcal{A}, P)$, a *random interval* is a Borel measurable mapping $X : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ w.r.t. the well-known Hausdorff metric on $\mathcal{K}_c(\mathbb{R})$ [11]. Equivalently, a mapping $X : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ is an RI if both $\mathrm{mid}\, X$ and $\mathrm{spr}\, X$ are (real-valued) random variables. It is clear that $\mathrm{spr}\, X \geq 0$.

The usual interval arithmetic is expressed in terms of the $(\mathrm{mid}\,, \mathrm{spr}\,)$ representation as follows:

$$A_1 + \lambda A_2 = [(\mathrm{mid}\, A_1 + \lambda \mathrm{mid}\, A_2) \pm (\mathrm{spr}\, A_1 + |\lambda| \mathrm{spr}\, A_2)], \tag{1}$$

for $A_1, A_2 \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. The expected value of an RI X is defined in terms of the Aumann expectation [1] and it fulfils that $E([\mathrm{mid}\, X \pm \mathrm{spr}\, X]) = [E(\mathrm{mid}\, X) \pm E(\mathrm{spr}\, X)]$, whenever $\mathrm{mid}\, X, \mathrm{spr}\, X \in L_1(\Omega, \mathcal{A}, P)$.

2.1 Similarity Degree between Intervals

Let $A \in \mathcal{K}_c(\mathbb{R})$. The Lebesgue measure of A is given by $\lambda(A) = 2\mathrm{spr}\, A$. In addition, the Lebesgue measure of the empty set is $\lambda(\emptyset) = 0$.

Let now $A, B \in \mathcal{K}_c(\mathbb{R})$. The Lebesgue measure of the intersection between A and B can be expressed as follows (cf. [14]):

$$\lambda(A \cap B) = \max\left\{0, \min\left\{2\mathrm{spr}\, A, 2\mathrm{spr}\, B, \mathrm{spr}\, A + \mathrm{spr}\, B - |\mathrm{mid}\, A - \mathrm{mid}\, B|\right\}\right\} \tag{2}$$

A measure of the degree of similarity between the intervals $A, B \in \mathcal{K}_c(\mathbb{R})$ can be defined, in accordance with the Jaccard coefficient [9], as

$$S(A, B) = \frac{\lambda(A \cap B)}{\lambda(A \cup B)}, \tag{3}$$

where $\lambda(A \cup B) = \lambda(A) + \lambda(B) - \lambda(A \cap B)$ and either A or B are assumed not to be reduced to a singleton. Clearly, $0 \leq S(A, B) \leq 1$ since $\lambda(A \cap B) \leq \lambda(A \cup B)$.

This measure satisfies that $S(A, B) = 0$ iff $A \cap B = \emptyset$, $S(A, B) = 1$ if $A \subset B$, and $S(A, B) \in (0, 1)$ iff $A \cap B \neq \emptyset$ and $A \neq B$.

In Figure 1 are gathered various possible similarity degrees.

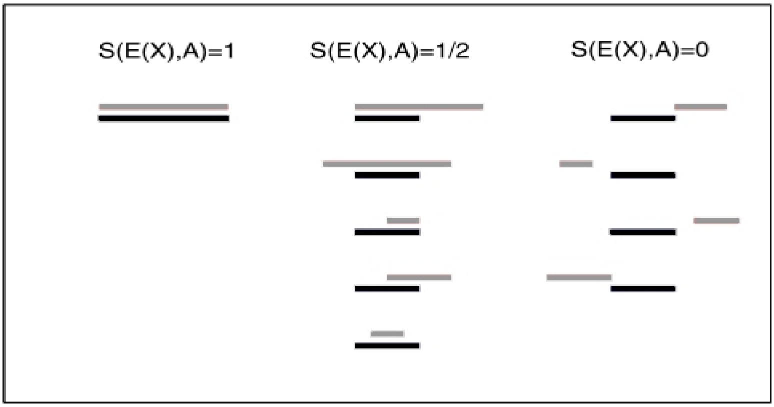


Fig. 1. Different representations for the similarity degree between $E(X)$ (in grey) and A (in black)

3 Hypothesis Testing for the Similarity between the Expected Value of an RI and a Fixed Interval

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $X : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ an RI so that $\operatorname{spr} E(X) > 0$ and $A \in \mathcal{K}_c(\mathbb{R})$ priory fixed so that $\operatorname{spr} A > 0$. Given $d \in [0,1]$, the aim is to test

$$H_0 : S(E(X), A) \geq d \text{ vs. } H_1 : S(E(X), A) < d. \tag{4}$$

The other one-sided test and the two-sided test could be analogously studied, but we will focus our attention in the previous one since it seems to be the most appealing for practical applications.

From the definition in (2), hypotheses of Test (4) can be equivalently expressed as

$$\begin{aligned} H_0 : \max\Big\{& d \operatorname{spr} A - \operatorname{spr} E(X), d \operatorname{spr} E(X) - \operatorname{spr} A, \\ & (1+d)\left|\operatorname{mid} E(X) - \operatorname{mid} A\right| + (d-1)\left(\operatorname{spr} E(X) + \operatorname{spr} A\right)\Big\} \leq 0; \\ H_1 : \max\Big\{& d \operatorname{spr} A - \operatorname{spr} E(X), d \operatorname{spr} E(X) - \operatorname{spr} A, \\ & (1+d)\left|\operatorname{mid} E(X) - \operatorname{mid} A\right| + (d-1)\left(\operatorname{spr} E(X) + \operatorname{spr} A\right)\Big\} > 0. \end{aligned} \tag{5}$$

3.1 Asymptotic Approach

If $\{X_i\}_{i=1}^n$ is a collection of random variables independent and identically distributed as X, the following test statistic is defined:

$$\begin{aligned} T_n = \sqrt{n} \max\Big\{& d \operatorname{spr} A - \operatorname{spr} \overline{X_n}, d \operatorname{spr} \overline{X_n} - \operatorname{spr} A, \\ & (1+d)\left|\operatorname{mid} \overline{X_n} - \operatorname{mid} A\right| + (d-1)\left(\operatorname{spr} \overline{X_n} + \operatorname{spr} A\right)\Big\}, \end{aligned} \tag{6}$$

where $\operatorname{mid}\overline{X_n}$ and $\operatorname{spr}\overline{X_n}$ are the corresponding classical sample means $\overline{\operatorname{mid} X_n}$ and $\overline{\operatorname{spr} X_n}$, respectively.

Some mild conditions are assumed to avoid trivial cases and to guarantee the existence of the involved moments. They are gathered in the following space:

$$\mathcal{P} = \left\{Y : \Omega \to \mathcal{K}_c(\mathbb{R}) \,|\sigma^2_{\operatorname{mid} Y} < \infty\,, 0 < \sigma^2_{\operatorname{spr} Y} < \infty \,\wedge\, \sigma^2_{\operatorname{mid} X,\, \operatorname{spr} X} \neq \sigma^2_{\operatorname{mid} X}\sigma^2_{\operatorname{spr} X}\right\}.$$

From now on, consider the bivariate normal distribution $Z = (z_1, z_2)^T \equiv \mathcal{N}_2\Big(\mathbf{0}, \Sigma\Big)$ where Σ is the covariance matrix for the random vector $(\operatorname{mid} X, \operatorname{spr} X)$. As we will show in the following lines, the limit distribution of the statistic T_n (and also the one of its bootstrap version) depend on the variables z_1 and z_2.

Lemma 1 shows the limit distribution of T_n under different conditions.

Lemma 1. *For $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be n RIs independent and equally distributed from X, and defined on the probability space $(\Omega, \mathcal{A}, P)$. Let T_n be defined as in (3.1). If $X \in \mathcal{P}$, then:*

a) *Whenever* $\operatorname{spr} E(X) = d \operatorname{spr} A$ *and* $\operatorname{mid} E(X) - \operatorname{mid} A = (1-d)\operatorname{spr} A$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{-z_2, (1+d)z_1 + (d-1)z_2\}. \tag{7}$$

b) *Whenever* $\operatorname{spr} E(X) = d \operatorname{spr} A$ *and* $-\operatorname{mid} E(X) + \operatorname{mid} A = (1-d)\operatorname{spr} A$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{-z_2, -(1+d)z_1 + (d-1)z_2\}. \tag{8}$$

c) *Whenever* $d \operatorname{spr} E(X) = \operatorname{spr} A$ *and* $\operatorname{mid} E(X) - \operatorname{mid} A = \dfrac{(1-d)}{d}\operatorname{spr} A$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{dz_2, (1+d)z_1 + (d-1)z_2\}. \tag{9}$$

d) *Whenever* $d \operatorname{spr} E(X) = \operatorname{spr} A$ *and* $-\operatorname{mid} E(X) + \operatorname{mid} A = \dfrac{(1-d)}{d}\operatorname{spr} A$, *it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{dz_2, -(1+d)z_1 + (d-1)z_2\}. \tag{10}$$

Proof. The statistic T_n can be equivalently expressed as:

$$\begin{aligned} T_n = \sqrt{n} \max \Big\{ & d \operatorname{spr} A - \operatorname{spr} E(X) + \operatorname{spr} E(X) - \operatorname{spr} \overline{X_n}, \\ & d \operatorname{spr} \overline{X_n} - d \operatorname{spr} E(X) + d \operatorname{spr} E(X) - \operatorname{spr} A, \\ & (1+d) \left|\operatorname{mid} \overline{X_n} - \operatorname{mid} E(X) + \operatorname{mid} E(X) - \operatorname{mid} A\right| \\ & +(d-1)\left(\operatorname{spr} \overline{X_n} - \operatorname{spr} E(X) + \operatorname{spr} E(X) + \operatorname{spr} A\right) \Big\}, \end{aligned}$$

a) If $\operatorname{spr} E(X) = d \operatorname{spr} A$ and $\operatorname{mid} E(X) - \operatorname{mid} A = (1-d)\operatorname{spr} A$, the second term and the negative form of the third term diverges in probability to $-\infty$ as $n \to \infty$ by the CLT and the Slutsky's theorem. Then, by using the continuous mapping theorem and the CLT for real variables (7) is provided. The same reasoning can be applied to the other three situations by taking into account that *b)* Whenever $\operatorname{spr} E(X) = d \operatorname{spr} A$ and $-\operatorname{mid} E(X) + \operatorname{mid} A = (1-d)\operatorname{spr} A$, the second term and the positive form of the third term diverges in probability to $-\infty$ as $n \to \infty$;

c) Whenever $d \operatorname{spr} E(X) = \operatorname{spr} A$ and $\operatorname{mid} E(X) - \operatorname{mid} A = \dfrac{(1-d)}{d}\operatorname{spr} A$, the first term and the negative form of the third term diverges in probability to $-\infty$ as $n \to \infty$;

d) Whenever $d \operatorname{spr} E(X) = \operatorname{spr} A$ and $-\operatorname{mid} E(X) + \operatorname{mid} A = \dfrac{(1-d)}{d}\operatorname{spr} A$, the first term and the negative form of the third term diverges in probability to $-\infty$ as $n \to \infty$.

□

Remark 1. It is easy to check that in other situations under H_0 the statistic T_n converges weakly to a limit distribution stochastically bounded for one of those provided in Lemma 1.

Since the limit distribution of T_n depends on X, it is suitable to consider the following X-dependent distribution for the theoretical analysis of the testing procedure:

$$\begin{aligned} T_n' = \max\Big\{ & \sqrt{n}\left(\operatorname{spr} E(X) - \operatorname{spr}\overline{X_n}\right) + \min\Big(0, n^{1/4}(\operatorname{spr} A - \operatorname{spr}\overline{X_n})\Big), \\ & \sqrt{n}\left(d\left(\operatorname{spr}\overline{X_n} - \operatorname{spr} E(X)\right)\right) + \min\Big(0, n^{1/4}(\operatorname{spr}\overline{X_n} - \operatorname{spr} A)\Big), \\ & \sqrt{n}\left((1+d)\left(\operatorname{mid}\overline{X_n} - \operatorname{mid} E(X)\right) + (d-1)\left(\operatorname{spr}\overline{X_n} - \operatorname{spr} E(X)\right)\right) \\ & + \min\Big(0, n^{1/4}(\operatorname{mid}\overline{X_n} - \operatorname{mid} A)\Big), \\ & \sqrt{n}\left((1+d)\left(\operatorname{mid} E(X) - \operatorname{mid}\overline{X_n}\right) + (d-1)\left(\operatorname{spr}\overline{X_n} - \operatorname{spr} E(X)\right)\right) \\ & + \min\Big(0, n^{1/4}(\operatorname{mid} A - \operatorname{mid}\overline{X_n})\Big) \Big\}. \end{aligned} \tag{11}$$

The minima included in T_n' is useful to determine the terms on its expression which has influence depending on the situation under H_0. For instance, if $\operatorname{spr} A - \operatorname{spr} E(X) \geq 0$ and $\operatorname{mid} A - \operatorname{mid} E(X) \geq 0$, the second and the third terms of T_n' diverge in probability to $-\infty$ whereas the first and the fourth terms determine the limit distribution of the statistic. Clearly, T_n' converge to the same distributions that T_n under the conditions established in Lemma 1. The consistency of the test is settled in the following lines.

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 (5) 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 Lemma 1 are fulfilled. In addition, if H_0 is not fulfilled then

$$\lim_{n\to\infty} P\left(T_n' > k_{1-\alpha}\right) = 1.$$

Therefore, the test which rejects H_0 in (5) at the significance level α whenever $T_n' > k_{1-\alpha}$ is asymptotically correct and consistent.

3.2 Bootstrap Approach

Due to the difficulties in handling the asymptotic limit distribution, a residual bootstrap approach is proposed.

Let X be an RI s.t. $\operatorname{spr} E(X) > 0$, $A \in \mathcal{K}_c(\mathbb{R})$ s.t. $\operatorname{spr} A > 0$ and $\{X_i\}_{i=1}^n$ be a simple random sample from X. Let $\{X_i^*\}_{i=1}^n$ be a bootstrap sample from $\{X_i\}_{i=1}^n$. The bootstrap statistic is defined below on the basis of T_n' and the classical residual bootstrap approach.

$$
\begin{aligned}
T_n^* = \max \Big\{ & \sqrt{n}\left(\operatorname{spr} \overline{X_n} - \operatorname{spr} \overline{X_n^*}\right) + \min\Big(0, n^{1/4}(\operatorname{spr} A - \operatorname{spr} \overline{X_n})\Big), \\
& \sqrt{n}\left(d\left(\operatorname{spr} \overline{X_n^*} - \operatorname{spr} \overline{X_n}\right)\right) + \min\Big(0, n^{1/4}(\operatorname{spr} \overline{X_n} - \operatorname{spr} A)\Big), \\
& \sqrt{n}\left((1+d)\left(\operatorname{mid} \overline{X_n^*} - \operatorname{mid} \overline{X_n}\right) + (d-1)\left(\operatorname{spr} \overline{X_n^*} - \operatorname{spr} \overline{X_n}\right)\right) \\
& + \min\Big(0, n^{1/4}(\operatorname{mid} \overline{X_n} - \operatorname{mid} A)\Big), \\
& \sqrt{n}\left((1+d)\left(\operatorname{mid} \overline{X_n} - \operatorname{mid} \overline{X_n^*}\right) + (d-1)\left(\operatorname{spr} \overline{X_n^*} - \operatorname{spr} \overline{X_n}\right)\right) \\
& + \min\Big(0, n^{1/4}(\operatorname{mid} A - \operatorname{mid} \overline{X_n})\Big) \Big\}.
\end{aligned}
\tag{12}
$$

The asymptotic distribution of T_n^* is provided in the following lemma.

Lemma 2. *Let $X \in \mathcal{P}$. Then,*

a) Whenever $\operatorname{spr} E(X) = d \operatorname{spr} A$ *and* $\operatorname{mid} E(X) - \operatorname{mid} A = (1-d)\operatorname{spr} A$*, it is fulfilled that*

$$T_n^* \xrightarrow{\mathcal{L}} \max\{-z_2, (1+d)z_1 + (d-1)z_2\} \;\; a.s. - [P]. \tag{13}$$

b) Whenever $\operatorname{spr} E(X) = d \operatorname{spr} A$ *and* $-\operatorname{mid} E(X) + \operatorname{mid} A = (1-d)\operatorname{spr} A$*, it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{-z_2, -(1+d)z_1 + (d-1)z_2\} \;\; a.s. - [P]. \tag{14}$$

c) Whenever $d \operatorname{spr} E(X) = \operatorname{spr} A$ *and* $\operatorname{mid} E(X) - \operatorname{mid} A = \dfrac{(1-d)}{d}\operatorname{spr} A$*, it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{d z_2, (1+d)z_1 + (d-1)z_2\} \;\; a.s. - [P]. \tag{15}$$

d) Whenever $d \operatorname{spr} E(X) = \operatorname{spr} A$ *and* $-\operatorname{mid} E(X) + \operatorname{mid} A = \dfrac{(1-d)}{d}\operatorname{spr} A$*, it is fulfilled that*

$$T_n \xrightarrow{\mathcal{L}} \max\{d z_2, -(1+d)z_1 + (d-1)z_2\} \;\; a.s. - [P]. \tag{16}$$

Remark 2. The consistency of the bootstrap procedure can be easily proven. In addition, other situations under H_0 leads to other limit distributions of the bootstrap statistic different from the ones provided in Lemma 2.

In practice, Monte Carlo method is employed to approximate the distribution of T_n^*.

4 Simulations

Some simulated models are proposed in order to analyze the behaviour of the bootstrap approach. Given the RI X, two different models are considered depending on the distributions for its mid and its spread, mainly,

Case 1: $\operatorname{mid} X \equiv \mathcal{N}(1,5)$ and $\operatorname{spr} X \equiv U(0,4)$;

Case 2: $\operatorname{mid} X \equiv U(-4,6)$ and $\operatorname{spr} X \equiv \chi^2_2$.

Let $A = [-5,3] \in \mathcal{K}_c(\mathbb{R})$. The aim is to construct the test

$$H_0 : S(E(X), A) \geq 1/2 \quad \text{vs.} \quad H_1 : S(E(X), A) < 1/2.$$

The bootstrap approach in Section 3.2 has been applied. Specifically, 10000 simulations with 1000 bootstrap replications have been carried out at the usual significance levels ρ (.01, .05 and .1) for different sample sizes. Results are gathered in Table 1.

Table 1. Empirical size of the bootstrap tests for the similarity degree

	Case 1			Case 2		
$n \diagdown 100 \cdot \rho$	1	5	10	1	5	10
10	2.82	7.94	13.66	3.63	8.11	11.40
30	1.59	5.74	10.88	1.80	5.81	11.08
50	1.35	5.36	10.59	1.67	5.66	10.92
100	1.27	5.26	10.35	1.28	5.32	10.26
200	1.10	5.06	10.08	1.09	5.02	10.10

Table 1 shows that the empirical sample sizes are in both cases quite close to the nominal significance levels for sample sizes greater than or equal to $n = 100$. However, the approximation to the nominal significance level is faster in Case 1 with respect to Case 2, especially for small sample sizes, which could be due to the differences in nature of the distributions involved in both models.

On the other hand, although the theoretical study of the power of the test will be developed in the future, some small simulations have been carried out. Specifically, $\operatorname{mid} X$ in Case 1 has been chosen to have distributions $\mathcal{N}(3,5)$, $\mathcal{N}(5,5)$ and $\mathcal{N}(7,5)$, respectively. In these cases, the bootstrap approach for $\alpha = .05$ and $n = 10$ lead to p-values of .39, .814 and .984, respectively, which implies that 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 test for analyzing the similarity between the expected value of an RI and a prefixed interval has been developed. Asymptotic and bootstrap techniques have been tackled and some simulations have been carried out showing the suitability of the bootstrap approach for moderate/large sample sizes.

In the future, a theoretical and empirical comparison between the Jaccard coefficient and other different similarity indexes could be established. In addition, different test statistics involving the covariance matrix can be studied as well as the influence of the distributions chosen for the simulations in order to reduce the bias observed for small sample sizes. The power of the test may be theoretically analyzed. Finally, it could be also interesting to extend the results provided in this work to the case of fuzzy sets.

References

1. Aumann, R.J.: Integrals of set-valued functions. Journal of Mathematical Analysis and Applications 12, 1–12 (1965)
2. Blanco-Fernández, A., Corral, N., González-Rodríguez, G.: Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput. Stat. Data An. 55(9), 2568–2578 (2011)
3. Diamond, P.: Least squares fitting of compact set-valued data. Journal of Mathematical Analysis and Applications 147, 531–544 (1990)
4. D'Urso, P., De Giovanni, L.: Midpoint radius self-organizing maps for interval-valued data with telecommunications application. Applied Soft Computing 11(5), 3877–3886 (2011)
5. Ferraro, M.B., Coppi, R., González-Rodríguez, G., Colubi, A.: A linear regression model for imprecise response. Int. J. Approx. Reason. 51(7), 759–770 (2010)
6. Gil, M.A., González-Rodríguez, G., Colubi, A., Montenegro, M.: Testing linear independence in linear models with interval-valued data. Computational Statistics & Data Analysis 51, 3002–3015 (2007)
7. Giordani, P., Kiers, H.A.L.: A comparison of three methods for principal component analysis of fuzzy interval data. Computational Statistics & Data Analysis 51, 379–397 (2006)
8. González-Rodríguez, G., Colubi, A., Gil, M.A.: Fuzzy data treated as functional data: A one-way ANOVA test approach. Computational Statistics and Data Analysis 56(4), 943–955 (2012)
9. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547–579 (1901)
10. Körner, R.: An asymptotic α-test for the expectation of random fuzzy variables. J. Stat. Plann. Inference 83, 331–346 (2000)
11. Matheron, G.: Random Sets and Integral Geometry. Wiley, New York (1975)
12. Molchanov, I.: Theory of Random Sets. Springer, London (2005)
13. Rivero, C., Valdes, T.: An algorithm for robust linear estimation with grouped data. Computational Statistics & Data Analysis 53, 255–271 (2008)
14. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
15. Sinova, B., Colubi, A., Gil, M.A., González-Rodríguez, G.: Interval arithmetic-based linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric. Inf. Sci. 199, 109–124 (2012)

Part III
Soft Methods in Data Analysis

Lasso Estimation of an Interval-Valued Multiple Regression Model

Marta García Bárzana[1], Ana Colubi[1], and Erricos John Kontoghiorghes[2]

[1] Department of Statistics. University of Oviedo,
C/ Calvo Sotelo s/n, Oviedo, 33007, Spain
[2] Department of Commerce, Finance and Shipping,
Cyprus University of Technology. P.O. Box 50329, CY-3603 Limassol, Cyprus
{garciabmarta,colubi}@uniovi.es, erricos@cut.ac.cy

Abstract. A multiple interval-valued linear regression model considering all the cross-relationships between the mids and spreads of the intervals has been introduced recently. A least-squares estimation of the regression parameters has been carried out by transforming a quadratic optimization problem with inequality constraints into a linear complementary problem and using Lemke's algorithm to solve it. Due to the irrelevance of certain cross-relationships, an alternative estimation process, the LASSO (Least Absolut Shrinkage and Selection Operator), is developed. A comparative study showing the differences between the proposed estimators is provided.

Keywords: Multiple regression, Lasso estimation, interval-valued data.

1 Introduction

Intervals represent a powerful tool to capture the imprecision of certain characteristics that cannot be fully described with a real number. For example, the measures provided by instruments which have some errors in their measurements [1]. Moreover, intervals also model some features which are inherently interval-valued. For instance, the range of variation of the blood preasure of a patient along a day [2] or the tidal fluctuation [9].

The statistical study of regression models for interval data has been extensively addressed lately in the literature [2–5, 7], deriving into several alternatives to tackle this problem. On one hand, the estimators proposed in [4, 7] account the non-negativity constraints satisfied by the spread variables, but do not assure the existence of the residuals. Hence, they can lead to ill-defined estimated models. On the other hand, the models proposed in [2, 3, 5] are formalized according to the natural interval arithmetic and their estimators lead to models that are always well-defined over the sample range.

The multiple linear regression model [3] considered belongs to the latter approach and its main advantage is the flexibility derived from its way to split the

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_22

regressors, allowing us to account for all the cross-relationships between the centers and the radius of the interval-valued variables. Nevertheless, this fact entails an increase in the number of regression parameters and thus, a Lasso estimation is considered in order to shrink some of these coefficients towards zero. The Lasso estimation of an interval-valued regression model has been previously addressed in [4], but it is a more restrictive model formalized in the first framework, where residuals might not exist.

The paper is organized as follows. Section 2 presents some preliminary concepts about the interval framework and section 3 contains the formalization of the model. The Least-Squares and Lasso estimations of the proposed model are developed in subsections 3.1 and 3.2. Section 4 briefly describes the Lasso model proposed by Giordani [4]. The empirical performance of the estimators proposed in sections 3 and 4 is compared in section 5 by means of a ilustrative real-life example. Section 6 finishes with some conclusions.

2 Preliminaries

Interval data are defined as elements belonging to the space $\mathcal{K}_{\mathrm{c}}(\mathbb{R}) = \{[a_1, a_2] : a_1, a_2 \in \mathbb{R}, a_1 \leq a_2\}$. Given an interval $A \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$, it can be parametrized in terms of its center or *midpoint*, $\operatorname{mid} A = (\sup A + \inf A)/2$, and its radius or *spread*, $\operatorname{spr} A = (\sup A - \inf A)/2$. Nonetheless, intervals can alternatively be expressed by means of the so-called *canonical decomposition* [2] defined as $A = \operatorname{mid}A[1 \pm 0] + \operatorname{spr}A[0 \pm 1]$, where $[1 \pm 0] = [1, 1]$ and $[0 \pm 1] = [-1, 1]$. This decomposition allows us to consider separately the *mid* and *spr* components of A, which will lead into a more flexible model. The interval arithmetic on $\mathcal{K}_{\mathrm{c}}(\mathbb{R})$ consists of the Minkowski addition and the product by scalars defined as follows by the jointly expression: $A + \delta B = [(\operatorname{mid}A + \delta \operatorname{mid}B) \ \pm \ (\operatorname{spr}A + |\delta| \operatorname{spr}B)]$ for any $A, B \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$ and $\delta \in \mathbb{R}$.

The space $(\mathcal{K}_{\mathrm{c}}(\mathbb{R}), +, \cdot\,)$ is not linear but semilinear, as the existence of symmetric element with respect to the addition is not guaranteed in general. An additional operation is introduced, the so-called Hukuhara difference between the intervals A and B. The difference C is defined as $C = A -_H B \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$ verifying that $A = B + C$. The existence of C is subject to the fulfilment of the expression $\operatorname{spr}B \leq \operatorname{spr}A$.

Given the intervals $A, B \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$, the metric $d_\tau(A, B) = ((1 - \tau)\,((\operatorname{mid}A - \operatorname{mid}B)^2 + \tau\,(\operatorname{spr}A - \operatorname{spr}B)^2))^{\frac{1}{2}}$, for an arbitrary $\tau \in (0, 1)$, is the L_2-type distance to be considered. d_τ is based on the metric d_θ defined in [11].

Given a probability space $(\Omega, \mathcal{A}, P)$ the mapping $\boldsymbol{x} : \Omega \to \mathcal{K}_{\mathrm{c}}(\mathbb{R})$ is a random interval iff it is a measurable Borel mapping. The moments to be considered are the classical Aumann expected value for intervals; the variance defined following the usual Fréchet variance [8] associated with the Aumann expectation in the interval space $(\mathcal{K}_{\mathrm{c}}(\mathbb{R}), d_\tau)$; and the covariance defined in terms of mids and spreads as $\sigma_{\boldsymbol{x},\boldsymbol{y}} = (1 - \tau)\,\sigma_{\operatorname{mid}\boldsymbol{x},\operatorname{mid}\boldsymbol{y}} + \tau\sigma_{\operatorname{spr}\boldsymbol{x},\operatorname{spr}\boldsymbol{y}}$.

3 The Multiple Linear Regression Model

Let $\boldsymbol{y}$ be a response random interval and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ be k explanatory random intervals. The model is formalized in a matrix notation as follows:

$$\boldsymbol{y} = X^{Bl}\, B + \boldsymbol{\varepsilon} \ , \tag{1}$$

where $B = (b_1|b_2|b_3|b_4)^t \in \mathbb{R}^{4k\times 1}$ with $b_i \in \mathbb{R}^k$ $(i \in \{1,2,3,4\})$, $X^{Bl} = (\boldsymbol{x^M}|\boldsymbol{x^S}|\boldsymbol{x^C}|\boldsymbol{x^R}) \in \mathcal{K}_c(\mathbb{R})^{1\times 4k}$ where the elements are defined as $\boldsymbol{x^M} = \operatorname{mid} x^t\,[1 \pm 0]$, $\boldsymbol{x^S} = \operatorname{spr} x^t\,[0 \pm 1]$, $\boldsymbol{x^C} = \operatorname{mid} x^t\,[0 \pm 1]$ and $\boldsymbol{x^R} = \operatorname{spr} x^t\,[1 \pm 0]$, considering the canonical decomposition of the regressors. Superscripts represent Bl=Block matrix, M=mid, S=spread, C=center and R=radius.

x^t is the vector of k explanatory random intervals, i.e., $x^t = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$. Thus, $\operatorname{mid} x^t = (\operatorname{mid} \boldsymbol{x_1}, \operatorname{mid} \boldsymbol{x_2}, \ldots, \operatorname{mid} \boldsymbol{x_k}) \in \mathbb{R}^k$ (analogously $\operatorname{spr} x^t$) and $\boldsymbol{\varepsilon}$ is a random interval-valued error such that $E(\boldsymbol{\varepsilon}|x) = \Delta \in \mathcal{K}_c(\mathbb{R})$.

The following separate linear relationships for the *mid* and *spr* components of the intervals are derived from (1):

$$\operatorname{mid} \boldsymbol{y} = \operatorname{mid} x^t\, b_1 + \operatorname{spr} x^t\, b_4 + \operatorname{mid} \boldsymbol{\varepsilon} \ , \tag{2a}$$

$$\operatorname{spr} \boldsymbol{y} = \operatorname{spr} x^t\, |b_2| + |\operatorname{mid} x^t|\, |b_3| + \operatorname{spr} \boldsymbol{\varepsilon} \ . \tag{2b}$$

Thus, the flexibility of the model arises from the possibility of considering all the information provided by mid$x$ and spr$x$ to model mid$\boldsymbol{y}$ and spr$\boldsymbol{y}$, as follows from (2a) and (2b). This represents an improvement with respect to previous models that merely addressed the relationship between the mids of the variables or between the spreads but never any cross-relationship (mid-spr).

Nevertheless, the inclusion of more coefficients entails an increase in the dimensionality of the estimation process. Some of these coefficients could be zero as not all the new introduced variables might contribute. Therefore it is proposed to estimate (1) by least-squares and by Lasso and compare the advantages and disadvantages that each estimation process provides.

3.1 The Least-Squares Estimation

Given $\{(\boldsymbol{y_j}, \boldsymbol{x_{i,j}}) : i = 1, \ldots, k, j = 1, \ldots, n\}$ a simple random sample of intervals obtained from $(\boldsymbol{y}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ in (1) the estimated model is

$$\widehat{y} = X^{ebl}\widehat{B} + \widehat{\varepsilon} \tag{3}$$

where $y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^t$, $X^{ebl} = (X^M|X^S|X^C|X^R) \in \mathcal{K}_c(\mathbb{R})^{n\times 4k}$ (the superscript *ebl* comes from estimated block matrix), $\varepsilon = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^t$ is such that $E(\varepsilon|x) = 1^n\Delta$ and B as in (1). X^M is the $(n \times k)$-interval-valued matrix such that $(X^M)_{j,i} = \operatorname{mid}\boldsymbol{x}_{i,j}[1 \pm 0]$ (analogously X^S, X^C and X^R). Given an arbitrary vector of regression coefficients $A \in \mathbb{R}^{4k\times 1}$ and an interval of residuals $C \in \mathcal{K}_c(\mathbb{R})$, the Least-Squares estimation looks for $\widehat{B}$ and $\widehat{\Delta}$ minimizing the distance $d_\tau^2(y, X^{ebl}A + 1^nC)$. $\widehat{\Delta}$ can be obtained separately and firstly by the expression $\widehat{\Delta} = \overline{\boldsymbol{y}} -_H \overline{X^{ebl}}\widehat{B}$.

Recalling that, by definition, $X^S = -X^S$ (and analogously $X^C = -X^C$) the estimation process of the coefficients b_2 and b_3 accompanying these variables can be simplified by searching only for non-negative estimates. By contrast, coefficients b_1 and b_4 are not affected by any kind of restrictions so they can be estimated directly by OLS. Moreover, it has to be assured the existence of the residuals defined as the Hukuhara differences $\varepsilon = \boldsymbol{y} -_H X^{ebl} B$. For this purpose the minimization problem ends up to be the following constrained quadratic problem:

$$\min_{A_m \in \mathbb{R}^{2k},\ A_s \in \Gamma} (1-\tau)\, \|v_m - F_m A_m\|^2 + \tau\, \|v_s - F_s A_s\|^2 \tag{4}$$

$$\Gamma = \{(a_2, a_3) \in [0,\infty)^k \times [0,\infty)^k : \operatorname{spr} X\, a_2 + |\operatorname{mid} X|\, a_3 \leq \operatorname{spr} y\},$$

being $v_m = \text{mid}y - \overline{\text{mid}\boldsymbol{y}}\, 1^n$, $v_s = \text{spr}y - \overline{\text{spr}\boldsymbol{y}}\, 1^n \in \mathbb{R}^n$, $F_m = \text{mid}X^{ebl} - 1^n(\overline{\text{mid}X^{ebl}})$, $F_s = \text{spr}X^{ebl} - 1^n(\overline{\text{spr}X^{ebl}}) \in \mathbb{R}^{n\times 2k}$, $A_m = (a_1|a_4)^t \in \mathbb{R}^{2k\times 1}$ the coefficients related to the midpoints and $A_s = (a_2|a_3)^t \in \mathbb{R}^{2k\times 1}$ the coefficients related to the spreads, with $a_l \in \mathbb{R}^k$, $l = 1, \ldots, 4$.

There are several numerical ways to tackle the resolution of a quadratic problem as (4). Given the shape of the objective function, the minimization process is solved separately over A_m and A_s. Those coefficients related with the mids (A_m) are not affected by constraints and therefore, the OLS estimator can be used directly. Thus $\widehat{A_m} = (F_m^t F_m)^{-1} F_m^t v_m$. However, in order to proceed with the constrained minimization over A_s, Karush-Kuhn-Tucker conditions guarantee the existence of local optima solution, which can be computed with standard numerical tool. Nevertheless, in order to obtain an exact solution and a more handy estimator of A_s, (4) can be equivalent expressed as a *Linear Complementary Problem* with the shape:

$$\omega = M\lambda + q \quad s.t. \quad \omega, \lambda \geq 0\ ,\ \omega_j \lambda_j = 0\ ,\ j = 1, \ldots, n+1\ , \tag{5}$$

with $M = (R\,Q^{-1}\,R^t)$ and $q = (-R\,Q^{-1}\,c - r)$ (details in [3]). Thereby, once λ is obtained, the expression of the estimator is $\widehat{A_s} = Q^{-1}\,(R^t\,\lambda - c)$.

3.2 The Lasso Estimation

Least Absolute Shrinkage and Selection Operator (LASSO) is a regression method that penalizes the sum of the absolute values of the regression coefficients estimates. For this purpose it involves a regularization parameter which affects directly the estimates: the larger the value of this parameter, the more estimates that are shrunk towards zero. However, this coefficient cannot be estimated statistically, so a cross-validation process is usually applied.

As previously, (4) can be solved separately. On one hand, the classical Lasso method will be used to obtain the estimator of the regression coefficients related to the mids. Thus, the problem is expressed as:

$$\frac{1}{2}\|v_m - A_m\, F_m\|_2^2 + \lambda \sum_{j=1}^{2k} |A_{m_j}|$$

being λ the regularization parameter. There are different programs capable to solve this problem (such as Matlab or R). The lasso.m Matlab function is the one used to obtain $\widehat{A_m}$.

On the other hand, for those coefficients related with the spreads a constrained Lasso algorithm has been developed as a modified version of the code proposed by Mark Schmidt (2005) [10] and is available upon request. The problem is given by:

$$\frac{1}{2}\|v_s - A_s\, F_s\|_2^2 + \lambda \sum_{j=1}^{2k} |A_{s_j}| \qquad \text{s.t } \; RA_s \geq r.$$

The most usual elections of λ are the value than minimizes the Cross-Validation Mean Square Error (λ_{MSE}) and the value that provides a simpler or more parsimonious model with respect to λ_{MSE} (in terms of more zero coefficients) but at the same time with one-standard-error (λ_{1SE}).

4 Giordani's Lasso Estimation

The so-called *Lasso-based Interval-valued Regression (Lasso-IR)* proposed by Giordani in [4] is another Lasso method to deal with a multiple linear regression model for interval data. However, the later regression model is not formalized following the interval arithmetic and can end up with an ill-defined estimated model. Keeping the same notation as in (2b), it requires the non-negativity of b_2 and b_3 but does not test if the Hukuhara's difference $\varepsilon = \boldsymbol{y} -_H X^{ebl} B$ exists. The optimization problem can be written (analogously to (4)) as:

$$\min_{A_m, As} (1-\tau)\,\|v_m - F_m\, A_m\|^2 + \tau \|v_s - F_s(A_m + A_a)\|^2 \tag{6}$$

$$F_s(A_m + A_a) \geq 0, \sum_{j=0}^{p} |A_{a_j}| \leq t$$

The coefficients related to the spreads (A_s) are the ones for the mids (A_m) plus a vector of additive coefficients (A_a) showing the distance that they are allowed to differ from A_m. In this case (6) has been expressed as a constrained quadratic problem, where there is a one-to-one correspondence between λ and t. The value of t that minimizes the cross-validation mean square error is the one considered. In order to solve the problem a stepwise algorithm based on [6] is proposed.

Another important difference, which entails less flexibility in the model, is the limitation of being able to study separately the relationships between the mids and the relationship between the spreads of the intervals but never any cross-relationship.

Remark 1. There is a particular case of model (1), the so-called Model M addressed in [2], which is formalized in the interval framework but has the same lack of flexibility as (6). In this case b_3 and $b_4 = (0, \ldots, 0)$, so the model has the shape:

$$\boldsymbol{y} = b_1\, \boldsymbol{x}^{\boldsymbol{M}} + b_2\, \boldsymbol{x}^{\boldsymbol{S}} + \boldsymbol{\varepsilon}. \tag{7}$$

5 A Real-Life Illustrative Example

The following example contains the information of a sample of 59 patients (from a population of 3000) hospitalized in the Hospital Valle del Nalón in Asturias, Spain. The variables to be considered are the ranges of fluctuation of the diastolic blood preasure over the day ($\boldsymbol{y}$), the pulse rate ($\boldsymbol{x_1}$) and the systolic blood preasure ($\boldsymbol{x_2}$). The dataset can be found in [2] and [5].

In order to make possible the comparison between the estimator proposed in section 4 and those ones introduced in subsections 3.1 and 3.2, the example will be developed for the simpler model explained in Remark 1.

Given the displayed model in (7), $\boldsymbol{y} = b_1\boldsymbol{x_1^M} + b_2\boldsymbol{x_2^M} + b_3\boldsymbol{x_1^S} + b_4\boldsymbol{x_2^S} + \varepsilon$, the estimates of the regression coefficients are summarized in Table 1:

Table 1. Estimates of the regression coefficients for the three estimators: LS, Lasso (for the two more representatives values of λ) and Lasso-IR (for a fixed value of t=0.10 prefixed by the author). The last column contains the MSE of the models mimicking its definition in the classical framework.

	$\widehat{b_1}$	$\widehat{b_2}$	$\widehat{b_3}$	$\widehat{b_4}$	MSE
$LS - estimation\,(Sect.\,3.1)$	0.4497	0.0517	0.2588	0.1685	68.2072
$Lasso - estimation\,(Sect.\,3.2)$	0.4202	0.0020	0.3379	0.2189	68.8477
λ_{MSE}	(0.6094)		(0.0259)		
$Lasso - estimation\,(Sect.\,3.2)$	0.2749	0	0.0815	0	76.9950
λ_{1SE}	(3.2521)		(1.8736)		
$Lasso - IR\,(Sect.4)$	0.5038	0.1261	0.4847	0.3605	71.2418

In view of the results in Table 1, those coefficients which take small values with the LS-estimation ($\widehat{b_2}$ and $\widehat{b_4}$) are schrunk towards zero with the most preferable Lasso estimation (for λ_{1SE}). However, this entails a significant increase of the MSE. In the case of using our Lasso-estimator for λ_{MSE}, the MSE is smaller but it does not provide a parsimonious model, being therefore its usefulness questionable. The estimator proposed in section 4 reaches a high value of MSE (worse in comparison with the lasso for λ_{MSE}) and does not end up with an easy-to-interpret model.

6 Conclusions

On one hand, a recently studied regression model for interval data, allowing to study all the cross-relationships between the mids and spreads of the interval-valued variables involved, is considered. This flexibility derives into an increase of the dimensionality of the model. Therefore a Lasso estimation seems appropiate to tackle this problem by setting some of these coefficients to zero. Nonetheless, a comparison study gathering the double estimation process conducted (first by Least-Squares and after by Lasso) is provided.

On the other hand, it is considered the Lasso-based interval-valued regression model (Lasso-IR) proposed in [4]. This model is not constrained to guarantee the existance of the residuals so it can provide misleading estimations. Moreover, it has a lack of flexibility as it solely tackles the relationships of type mid-mid and spr-spr but no cross-relationships mid-spr.

A real-life example illustrating the difference between the estimators in terms of MSE and simplicity has been conducted.

Acknowledgments. Financial support from the Spanish Ministerio de Ciencia e Innovacion (MICINN) through Ayuda Puente SV-PA-13-ECOEMP-66 and Acción Integrada PRI-AIBDE-2011-1197, is kindly acknowledged.

References

1. Abdallah, F., Gning, A., Bonnifait, P.: Adapting particle filter on interval data for dynamic state estimation. In: ICASSP, pp. 1153–1156 (2007)
2. Blanco-Fernández, A., Corral, N., González-Rodríguez, G.: Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput. Stat. Data Anals. 55, 2568–2578 (2011)
3. Blanco-Fernández, A., García-Bárzana, M., Colubi, A., Kontoghiorghes, E.J.: Multiple set arithmetic-based linear regression models for interval-valued variables (submitted)
4. Giordani, P.: Linear regression analysis for interval-valued data based on the Lasso technique. Adv. Data Anal. Classif. (2014) ISSN: 1862-5347
5. González-Rodríguez, G., Blanco, A., Corral, N., Colubi, A.: Least squares estimation of linear regression models for convex compact random sets. Adv. Data Anal. Classif. 1, 67–81 (2007)
6. Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. In: Classics in Applied Mathematics, vol. 15. SIAM, Philadelphia (1995)
7. Lima Neto, E.A., de Carvalho, F.A.T.: Constrained linear regression models for symbolic interval-valued variables. Comput. Stat. Data Anals. 54, 333–347 (2010)
8. Näther, W.: Linear statistical inference for random fuzzy data. Statistics 29(3), 221–240 (1997)
9. Ramos-Guajardo, A.B., González-Rodríguez, G.: Testing the Variability of Interval Data: An Application to Tidal Fluctuation. In: Borgelt, C., Gil, M.Á., Sousa, J.M.C., Verleysen, M. (eds.) Towards Advanced Data Analysis. STUDFUZZ, vol. 285, pp. 65–74. Springer, Heidelberg (2012)
10. Schmidt Mark, `http://www.di.ens.fr/~mschmidt`
11. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inf. Sci. 179(23), 3964–3972 (2009)

A Bootstrap Factorial ANOVA for Random Intervals

Angela Blanco-Fernández[1] and T. Warren Liao[2]

[1] Department of Statistics and O.R., University of Oviedo,
Calvo Sotelo s/n, Oviedo, 33007, Spain
[2] Department of Mechanical and Industrial Engineering,
Louisiana State University, Baton Rouge, LA 70803, U.S.A.
blancoangela@uniovi.es, ieliao@lsu.edu

Abstract. An ANOVA problem for interval-valued experimental data is considered. When a random variable is observed on several populations, the ANOVA technique focuses on testing whether the variable behaves significantly different on those groups. The theoretical formalization of the three-way ANOVA problem when the random element takes on interval values is shown. Since no distribution assumptions for interval-valued variables are established, a bootstrap technique for the statistical resolution of the inferential study is developed and implemented. The theoretical validity of the procedure is guaranteed from previous results, and its empirical behaviour is shown with a case study.

Keywords: interval-valued data, ANOVA problem, bootstrap test.

1 Introduction

The Analysis of Variance (ANOVA) problem is a well-known statistical technique to apply for real-valued random variables on a factorial design. When a real random variable is observed on different populations, established by the levels of some factors, it is often interesting to check if the variable has a similar behaviour on all the groups, or it performs significantly different depending on either the effect of one of the factors or the interaction between two or more factors.

Some classical statistical methods have been previously extended to deal with more general experimental scenarios, in which the experimental outcomes are no longer real numbers, but intervals, sets, or fuzzy values. In particular, real compact intervals are useful to represent experimental outcomes modelling fluctuations, grouped data, ranges of variation of a magnitude over a period of time or in a cross-section, interval-valued perceptions, among other examples [2, 14]. Additionally, intervals are also an effective tool for modelling experimental settings for which the outcomes can be measured only with some imprecision, which is reflected by a real interval rather than by a point-value [1].

Interval-valued random variables are treated in some works in the context of fuzzy- and set-valued models; see, for instance, [3, 4, 10]. Nevertheless, a number of methods has been developed for the interval framework specifically; see [2, 5–7, 12, 14], among others.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_23

It is important to remark that the approach in this work only considers the *imprecision* on the experimental data, but not on the statistical methods to manage those data, i.e. it follows the line of research of works developing classical statistical techniques with *imprecise-valued* data.

The Analysis of the Variance has been extended to this experimental scenarios [8, 9, 11–13]. In [8] a one-way ANOVA test for fuzzy data is proposed, which is based on functional data analysis. In [9] a bootstrap testing algorithm to solve that problem included in the R package SAFD is described. Those previous works are extended in [13] to the factorial ANOVA for fuzzy data. Specifically for intervals, in [12] a two-way ANOVA problem is established and a bootstrap test is developed theoretically. The aim of this paper is to formulate the three-way analysis of variance for interval-valued variables and to apply the technique to a case study. The procedure examines the significance of the effect of three different factors as well as all the possible interactions between the factors on the values of an interval-valued response. Theoretical results on [13] will support the validity of the method. Besides, an alternative bootstrap process is proposed, by following some ideas from [8], which makes easier the practical application of the technique.

The rest of the paper is organized as follows: in Section 2 some previous concepts and notions in the interval framework are recalled. Section 3 is devoted to the formalization of the three-way ANOVA for random intervals and the adaptation of a bootstrap test method to solve the problem. In Section 4 the empirical performance of the technique is investigated with the practical application in a case study. Finally, Section 5 includes some conclusions and future directions.

2 Preliminary Concepts

The space of real compact intervals is generally denoted as $\mathcal{K}_c(\mathbb{R})$. Each element of this space can be represented in terms of its end-points as $A = [\inf A, \sup A]$, with $\inf A \leq \sup A$, or, equivalently, in terms of its centre (midpoint) and its radious (spread) as $A = [\text{mid}A \pm \text{spr}A]$, being $\text{spr}A \geq 0$. In the statistical treatment of intervals it is usually employed the latter representation, since the non-negativity condition for the spreads is easier to handle in computations than the order condition between the end-points, in general. The natural arithmetic between intervals is composed on the so-called Minkowski addition and the product by scalars, defined as follows:

$$A + B = \{a + b : a \in A, b \in B\} \quad \text{and} \quad \lambda A = \{\lambda a : a \in A\} \ , \tag{1}$$

for any $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. These operations can be more intuitively expressed in terms of the (mid,spr)-representation of intervals as follows:

$$A + B = [(\text{mid}A + \text{mid}B) \pm (\text{spr}A + \text{spr}B)] \quad \text{and} \quad \lambda A = [\lambda\text{mid}A \pm |\lambda|\text{spr}A] \ . \tag{2}$$

It is straightforward to see that they are inner and natural operations in the space of intervals. Nevertheless, the arithmetic is not linear, but semilinear, due to the lack of symmetric element with respect to the addition, in general. Thus, statistical techniques in this space must be always developed by guaranteeing the coherency of the results with the semilinear structure of $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$.

In order to measure distances between intervals, a metric is defined on the space of intervals. Among the different alternatives existing in the literature, it is considered in this work a family of generalized L_2-type metrics exhaustively used in statistical developments for fuzzy-valued data (see [15]). For intervals it can be expressed as follows:

$$d_\tau(A, B) = \sqrt{(1-\tau)(\mathrm{mid}A - \mathrm{mid}B)^2 + \tau(\mathrm{spr}A - \mathrm{spr}B)^2} \,, \tag{3}$$

for certain $\tau \in (0, 1)$. The value of τ allows us to choose the relative importance of the squared Euclidean distance between the spreads of the intervals (difference in imprecision) and the squared Euclidean distance for the midpoints (difference in location).

The elements of the space $\mathcal{K}_c(\mathbb{R})$ will represent the values of a variable modelling a characteristic whose possible experimental outcomes are interval-valued. Thus, given $(\Omega, \mathcal{A}, P)$ a probability space, an interval-valued random variable (or random interval for short) is a mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ being $\mathcal{A}|\mathcal{B}_{d_\tau}$-measurable. The execution of the data generation process provides a simple random sample of intervals independent and equally distributed to X, $\{X_i\}_{i=1}^n$. The expected value of X, as the usual summary measure for central tendency, is generally defined in terms of the Aumann's expectation for set-valued functions; see [10]. For intervals, it admits the expression $E(X) = [E(\mathrm{mid}X) \pm E(\mathrm{spr}X)]$, whenever those classical moments exist. The sample counterpart is defined coherently with the interval arithmetic as $\overline{X} = (1/n)\sum_{i=1}^n X_i$.

3 Three-Way ANOVA for Random Intervals

The ANOVA problem for an interval-valued response variable on a 3^m-factorial design, i.e. the existence of three factors with m levels each, is formulated in this section. A bootstrap testing method is conducted to solve the problem.

3.1 Formulation of the Problem

Let X be a random interval which is observed under three factors F_1, F_2 and F_3, with I_1, I_2 and I_3 levels, respectively. Let $X_{i_1,i_2,i_3,k}$ denote the kth-observation of X under level i_j of factor F_j, $j = 1, 2, 3$. For the sake of simplicity on the notation and computations, let us assume a balanced design, i.e. that the number of observations for each group of levels (i_1, i_2, i_3) is equal to $n \in \mathbb{N}$, and so $k = 1, \ldots, n$ in all the groups. The total number of observations is then $n_T = nI_1I_2I_3$.

Each interval-valued (random) element $X_{i_1,i_2,i_3,k}$ can be modelled in terms of the interval arithmetic as follows [13]:

$$X_{i_1,i_2,i_3,k} = M + A_{i_1} + B_{i_2} + C_{i_3} + G_{i_1,i_2} + G_{i_1,i_3} + G_{i_2,i_3} + G_{i_1,i_2,i_3} + \varepsilon_{i_1,i_2,i_3,k} \,, \tag{4}$$

for all $i_j = 1, \ldots, I_j$, $j = 1, 2, 3$, where all the addends are intervals, representing:

- the overall baseline level: M. It is the component of $X_{i_1,i_2,i_3,k}$ which is not affected by any of the factors nor the interactions between them,
- the possible effect of each factor: A_{i_1}, B_{i_2}, C_{i_3}, respectively,
- the possible effect of the interactions of two factors: G_{i_1,i_2}, G_{i_1,i_3}, G_{i_2,i_3},
- the possible effect of the interactions of the three factors: G_{i_1,i_2,i_3}, and
- the error term: $\varepsilon_{i_1,i_2,i_3,k}$. It is the random component of the model. No distribution assumptions are established for the interval-valued error.

The aim is to determine the significance of the effect of the factors on X individually as well as by means of the interaction between two or the three of them. The following hypothesis tests are formulated to solve this problem:

- Test (1): $H_0^{(1)} : A_1 = A_2 = ... = A_{I_1}$ vs. $H_1^{(1)} : \exists i_{j_1}, i_{j_2}$ such that $A_{i_{j_1}} \neq A_{i_{j_2}}$.
- Test (2): $H_0^{(2)} : B_1 = B_2 = ... = B_{I_2}$ vs. $H_1^{(2)} : \exists i_{j_1}, i_{j_2}$ such that $B_{i_{j_1}} \neq B_{i_{j_2}}$.
- Test (3): $H_0^{(3)} : C_1 = C_2 = ... = C_{I_1}$ vs. $H_1^{(3)} : \exists i_{j_1}, i_{j_2}$ such that $C_{i_{j_1}} \neq C_{i_{j_2}}$.
- Test (1,2):
  $H_0^{(1,2)} : G_{i_{j_1},i_{j_2}} = G_{i'_{j_1},i'_{j_2}}$ for all $i_{j_1}, i'_{j_1} = 1, \ldots, I_1$ and $i_{j_2}, i'_{j_2} = 1, \ldots, I_2$ vs.
  $H_1^{(1,2)} : \exists i_{j_1}, i'_{j_1} \in 1, \ldots, I_1$ and $i_{j_2}, i'_{j_2} \in \{1, \ldots, I_2\}$ s.t. $G_{i_{j_1},i_{j_2}} \neq G_{i'_{j_1},i'_{j_2}}$.
- Test (1,3):
  $H_0^{(1,2)} : G_{i_{j_1},i_{j_3}} = G_{i'_{j_1},i'_{j_3}}$ for all $i_{j_1}, i'_{j_1} = 1, \ldots, I_1$ and $i_{j_3}, i'_{j_3} = 1, \ldots, I_3$ vs.
  $H_1^{(1,3)} : \exists i_{j_1}, i'_{j_1} \in 1, \ldots, I_1$ and $i_{j_3}, i'_{j_3} \in \{1, \ldots, I_2\}$ s.t. $G_{i_{j_1},i_{j_3}} \neq G_{i'_{j_1},i'_{j_3}}$.
- Test (2,3):
  $H_0^{(2,3)} : G_{i_{j_2},i_{j_3}} = G_{i'_{j_2},i'_{j_3}}$ for all $i_{j_2}, i'_{j_3} = 1, \ldots, I_2$ and $i_{j_3}, i'_{j_3} = 1, \ldots, I_3$ vs.
  $H_1^{(2,3)} : \exists i_{j_2}, i'_{j_2} \in 1, \ldots, I_2$ and $i_{j_3}, i'_{j_3} \in \{1, \ldots, I_3\}$ s.t. $G_{i_{j_2},i_{j_3}} \neq G_{i'_{j_2},i'_{j_3}}$.
- Test (1,2,3):
  $H_0^{(1,2,3)} : G_{i_{j_1},i_{j_2},i_{j_3}} = G_{i'_{j_1},i'_{j_2},i'_{j_3}}$ for all $i_{j_1}, i'_{j_1} = 1, \ldots, I_1$, $i_{j_2}, i'_{j_2} = 1, \ldots, I_2$ and $i_{j_3}, i'_{j_3} = 1, \ldots, I_3$ vs.
  $H_1^{(1,2,3)} : \exists i_{j_1}, i'_{j_1} \in 1, \ldots, I_1$, $i_{j_2}, i'_{j_2} \in \{1, \ldots, I_2\}$ and $i_{j_3}, i'_{j_3} = 1, \ldots, I_3$ s.t. $G_{i_{j_1},i_{j_2},i_{j_3}} \neq G_{i'_{j_1},i'_{j_2},i'_{j_3}}$.

Tests (1), (2) and (3) study the effect of each factor F_1, F_2 and F_3 individually on X, respectively. Tests (1,2), (1,3) and (2,3) study the effect of the interaction between the two corresponding factors. Finally, test (1,2,3) studies the effect of the interaction of the three factors on X.

3.2 Bootstrap Testing Method

The resolution of the preceding tests is developed by following a bootstrap approach. The process does not require the usual assumptions of classical ANOVA techniques on the distribution of the variable. Due to the semilinear structure of the space of intervals, the test statistics are defined in terms of distances (through the d_τ-metric) between sample means of intervals in certain groups of levels [12, 14]. Let us define first those sample means:

- $\overline{X_{\dots\cdot}} = \frac{1}{n_T} \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \sum_{k=1}^{n} X_{i_1,i_2,i_3,k}$: sample mean of all the interval observations, in all the groups.
- $\overline{X_{i_1\dots}} = \frac{1}{nI_2I_3} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \sum_{k=1}^{n} X_{i_1,i_2,i_3,k}$: sample mean of X in the level i_1 of the factor F_1, for each $i_1 = 1, \dots, I_1$.

  (analogously $\overline{X_{\cdot i_2 \cdot\cdot}}$ and $\overline{X_{\cdot\cdot i_3 \cdot}}$)
- $\overline{X_{i_1,i_2\cdot\cdot}} = \frac{1}{nI_3} \sum_{i_3=1}^{I_3} \sum_{k=1}^{n} X_{i_1,i_2,i_3,k}$: sample mean of X in the levels i_1 of the factor F_1 and i_2 of the factor F_2, for each $i_1 = 1, \dots, I_1$ and $i_2 = 1, \dots, I_2$.

  (analogously $\overline{X_{i_1 \cdot i_3 \cdot}}$ and $\overline{X_{\cdot i_2, i_3 \cdot}}$)
- $\overline{X_{i_1,i_2,i_3\cdot}} = \frac{1}{n} \sum_{k=1}^{n} X_{i_1,i_2,i_3,k}$: sample mean of X in the group of levels (i_1, i_2, i_3) of the factors F_1, F_2 and F_3, respectively, for each $i_1 = 1, \dots, I_1$, $i_2 = 1, \dots, I_2$ and $i_3 = 1, \dots, I_3$.

Let us consider the test (1). The statistic to test $H_0^{(1)}$ *vs.* $H_1^{(1)}$ is defined as

$$T_n^{(1)} = nI_2I_3 \sum_{i_1}^{I_1} d_\tau^2(\overline{X_{i_1\dots}}, \overline{X_{\dots\cdot}}) \,. \tag{5}$$

If $H_0^{(1)}$ is true, the sample means on all the levels i_1 of F_1 are equal each other and they also equal the global mean $\overline{X_{\dots\cdot}}$, so that $T_n^{(1)} = 0$. On the contrary, under $H_1^{(1)}$, $\overline{X_{i_1\dots}}$ are different from $\overline{X_{\dots\cdot}}$ and so $T_n^{(1)} > 0$. As a conclusion, the null hypothesis $H_0^{(1)}$ is rejected for large values of $T_n^{(1)}$.

The limit distribution of $T_n^{(1)}$ can be obtained [13]. However, that distribution is usually unknown in practice, so a bootstrap technique is applied, which does not require the distribution function of the statistic to be known. Following the bootstrap scheme proposed in [8], the bootstrap algorithm to test $H_0^{(1)}$ is designed as follows:

Bootstrap Algorithm to Test the Significance of the Factor $\mathbf{F_1}$ on X

1. Compute the sample means $\overline{X_{i_1\dots}}$, for $i_1 = 1, \dots, I_1$, and $\overline{X_{\dots\cdot}}$, and the value of the test statistic

$$T_n^{(1)} = nI_2I_3 \sum_{i_1}^{I_1} d_\tau^2(\overline{X_{i_1\dots}}, \overline{X_{\dots\cdot}}) \,. \tag{6}$$

2. For each group of levels (i_1, i_2, i_3) of factors F_1, F_2 and F_3, respectively, consider the sample of intervals on this group $\{X_{i_1,i_2,i_3,k}\}_{k=1}^n$, and generate

a bootstrap sample on this group $\{X^*_{i_1,i_2,i_3,k}\}_{k=1}^n$ by re-sampling randomly and with replacement n intervals from the preceding set. Thus, the complete bootstrap sample of X is

$$\bigcup_{i_1=1}^{I_1} \bigcup_{i_2=1}^{I_2} \bigcup_{i_3=1}^{I_3} \{X^*_{i_1,i_2,i_3,k}\}_{k=1}^n \, . \tag{7}$$

3. Based on this bootstrap sample, compute the corresponding sample means of X, $\overline{X^*_{i_1\cdot\cdot\cdot}}$, for $i_1 = 1, \ldots, I_1$, and $\overline{X^*_{\cdot\cdot\cdot\cdot}}$. The bootstrap test statistic is then computed as

$$T_n^{(1)*} = nI_2I_3 \sum_{i_1}^{I_1} d_\tau^2(\overline{X^*_{i_1\cdot\cdot\cdot}} + \overline{X_{\cdot\cdot\cdot\cdot}}, \overline{X^*_{\cdot\cdot\cdot\cdot}} + \overline{X_{i_1\cdot\cdot\cdot}}) \, . \tag{8}$$

4. Repeat Steps 2 and 3 a large number B of times. Approximate the p-value of the test by the proportion of values in $\{T_{n,b}^{(1)*}\}_{b=1}^B$ being greater than $T_n^{(1)}$.

It can be shown that the empirical distribution of the bootstrap test statistic $\{T_{n,b}^{(1)*}\}_{b=1}^B$ approximates the distribution of $T_n^{(1)}$ under $H_0^{(1)}$ [8].

Analogous algorithms are designed to solve the remainder hypothesis tests of the ANOVA problem. The differences in the algorithms appear in the definition of the statistic for each test in Eq. (6) as well as in the corresponding bootstrap statistic in Eq. (8). $T_n^{(2)}$ and $T_n^{(3)}$ (and bootstrap counterparts) are analogous to $T_n^{(1)}$. To solve the test $(1,2)$, we define in (6) and (8), respectively:

$$T_n^{(1,2)} = nI_3 \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} d_\tau^2(\overline{X_{i_1,i_2\cdot\cdot}} + \overline{X_{\cdot\cdot\cdot\cdot}}, \overline{X_{i_1\cdot\cdot\cdot}} + \overline{X_{\cdot i_2\cdot\cdot}}) \text{ , and}$$

$$T_n^{(1,2)*} = nI_3 \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} d_\tau^2\Big((\overline{X^*_{i_1,i_2\cdot\cdot}} + \overline{X^*_{\cdot\cdot\cdot\cdot}} + \overline{X_{i_1\cdot\cdot\cdot}} + \overline{X_{\cdot i_2\cdot\cdot}}), (\overline{X^*_{i_1\cdot\cdot\cdot}} + \overline{X^*_{\cdot i_2\cdot\cdot}} + \overline{X_{i_1,i_2\cdot\cdot}} + \overline{X_{\cdot\cdot\cdot\cdot}})\Big).$$

Analogously for tests $(1,3)$ and $(2,3)$. Finally, test $(1,2,3)$ is solved by defining the test statistics in (6) and (8) as

$$\begin{aligned} T_n^{(1,2,3)} = n \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} d_\tau^2\Big(& (\overline{X_{i_1,i_2,i_3\cdot}} + \overline{X_{i_1\cdot\cdot\cdot}} + \overline{X_{\cdot i_2\cdot\cdot}} + \overline{X_{\cdot\cdot i_3\cdot}}), \\ & (\overline{X_{i_1,i_2\cdot\cdot}} + \overline{X_{i_1\cdot i_3\cdot}} + \overline{X_{\cdot i_2,i_3\cdot}} + \overline{X_{\cdot\cdot\cdot\cdot}})\Big) \text{ , and} \end{aligned}$$

$$\begin{aligned} T_n^{(1,2,3)*} = n \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} d_\tau^2\Big(& (\overline{X^*_{i_1,i_2,i_3\cdot}} + \overline{X^*_{i_1\cdot\cdot\cdot}} + \overline{X^*_{\cdot i_2\cdot\cdot}} + \overline{X^*_{\cdot\cdot i_3\cdot}}) + \\ & (\overline{X_{i_1,i_2\cdot\cdot}} + \overline{X_{i_1\cdot i_3\cdot}} + \overline{X_{\cdot i_2,i_3\cdot}} + \overline{X_{\cdot\cdot\cdot\cdot}}), \\ & (\overline{X^*_{i_1,i_2\cdot\cdot}} + \overline{X^*_{i_1\cdot i_3\cdot}} + \overline{X^*_{\cdot i_2,i_3\cdot}} + \overline{X^*_{\cdot\cdot\cdot\cdot}}) + \\ & (\overline{X_{i_1,i_2,i_3\cdot}} + \overline{X_{i_1\cdot\cdot\cdot}} + \overline{X_{\cdot i_2\cdot\cdot}} + \overline{X_{\cdot\cdot i_3\cdot}})\Big) \, . \end{aligned}$$

4 Application in a Case Study

A case study about a sequencing problem of inbound trucks in a multi-door cross docking system is investigated. The experimental data set follow a 3^3-factorial design, where the response variable $X =$*makespan of a truck in the cross docking terminal* is interval-valued, and the factors $b \in \{1, 2, 3\}$, $\rho \in \{0.1, 0.2, 0.3\}$ and $\phi \in \{0.1, 0.2, 0.3\}$ determine the possible values of three algorithmic parameters of an ant colony optimization metaheuristic for solving the sequencing problem.

Let us denote $F_1 = b$, $F_2 = \rho$ and $F_3 = \phi$. $I_j = 3$, for all $j = 1, 2, 3$. The data set fulfils $n = 5$ and so $n_T = 135$[1].

The aim is to test the significance of the effect of each factor on X, as well as the significance of the effect of any of the possible interactions between the factors. Let $\tau = .25$ (see [15] for details). The proposed bootstrap algorithms are run for $B = 5000$ bootstrap iterations. The results are shown in Table 1.

Table 1. Bootstrap p-values for the ANOVA hypothesis tests

Test	p-value
(1,2,3)	.3978
(1,2)	.1755
(1,3)	.4434
(2,3)	.7222
(1)	0
(2)	0
(3)	.2356

As a conclusion, none of the interactions between factors affect significantly the response. Individually, the effect of the factor $F_3 = \phi$ on X is not significant neither. On the contrary, it is obtained that the factors $F_1 = b$ and $F_2 = \rho$ do affect significantly the response, i.e. the makespan behaves statistically different for the possible values of the parameters b and ρ in the sequencing problem.

5 Conclusions and Future Directions

In this work, the three-way ANOVA for interval-valued random variables is formalized. Hypothesis tests to check the effects of the main factors and the interactions between factors are formulated, and bootstrap algorithms to solve these tests are designed. Besides the theoretical validity of the procedure, the test algorithms have been implemented, making possible the application of the ANOVA problem in practice. Simulation studies are to be done. A multiple significance testing of the group of hypotheses by taking into account a kind of correction for the significance level could also be developed in future research.

The extension to the general m-way ANOVA problem, i.e. to consider the existence of $m \in \mathbb{N}$ factors, for an interval-valued response is theoretically straightforward, by introducing a heavy (but unavoidable) notation for the factorial

[1] The data set is available upon request to the authors.

model and the corresponding sample means and test statistics [13]. This heavy formulation implies that the implementation of the general m-way ANOVA for intervals scheme becomes hard, and it has not been developed yet.

References

1. Abdallah, F., Gning, A., Bonnifait, P.: Adapting particle filter on interval data for dynamic state estimation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol. 2, pp. 1153–1156 (2007)
2. Blanco-Fernández, A., Corral, N., González-Rodríguez, G.: Estimation of a flexible simple linear model for interval data based on set arithmetic. Computational Statistics & Data Analysis 55(9), 2568–2578 (2011)
3. Diamond, P.: Least squares fitting of compact set-valued data. Journal of Mathematical Analysis and Applications 147, 531–544 (1990)
4. D'Urso, P.: Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Computational Statistics & Data Analysis 42, 47–72 (2003)
5. D'Urso, P., Giordani, P.: A robust fuzzy k-Means clustering model for interval valued data. Computational Statistics 21, 251–269 (2006)
6. D'Urso, P., De Giovanni, L., Massari, R.: Trimmed fuzzy clustering for interval-valued data. Advances Data Analysis and Classification (in press, 2014)
7. Gil, M.A., López-García, M.T., Lubiano, M.A., Montenegro, M.: Regression and correlation analyses of a linear relation between random intervals. Test 10, 183–201 (2001)
8. González-Rodríguez, G., Colubi, A., Gil, M.A.: Fuzzy data treated as functional data: A one-way ANOVA test approach. Computational Statistics and Data Analysis 56, 943–955 (2012)
9. Lubiano, M.A., Trutschnig, W.: ANOVA for fuzzy random variables using the R-package SAFD. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano, M.A., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. AISC, vol. 77, pp. 449–456. Springer, Heidelberg (2010)
10. Molchanov, I.: Theory of random sets. Probability and its aplications. Springer-Verlag, London (2005)
11. Montenegro, M., González-Rodríguez, G., Gil, M.A., Colubi, A., Casals, M.R.: Introduction to ANOVA with fuzzy random variables. In: López-Díaz, et al. (eds.) Soft Methodology and Random Information Systems. AISC, vol. 26, pp. 487–494. Springer, Heidelberg (2004)
12. Nakama, T., Colubi, A., Lubiano, M.A.: Two-way analysis of variance for interval-valued data. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano, M.A., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. AISC, vol. 77, pp. 475–482. Springer, Heidelberg (2010)
13. Nakama, T.: Factorial analysis of variance for fuzzy data. Master Project, University of Oviedo (2010)
14. Ramos-Guajardo, A.B., González-Rodríguez, G.: Testing the variability of interval data: An application to tidal fluctuation. In: Borgelt, C., Gil, M.Á., Sousa, J.M.C., Verleysen, M., et al. (eds.) Towards Advanced Data Analysis. STUDFUZZ, vol. 285, pp. 65–74. Springer, Heidelberg (2012)
15. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences 179, 3964–3972 (2009)

Estimation of a Simple Multivariate Linear Model for Fuzzy Random Sets

Dabuxilatu Wang and Miao Shi

School of Economics and Statistics
Research Center of Statistical Science Lingnan,
Guangzhou University, No.230 Waihuanxi road,
Higher Education Mega Center, Guangzhou, 510006, P.R. China

Abstract. A simple two-variate linear regression model with fuzzy random sets under concepts of functional data analysis is considered. The support function of a fuzzy random set establishes a useful embedding of the space of fuzzy random sets into a cone of a functional Hilbert space. Treating the fuzzy random sets as special functional data, we estimate the linear model within the cone. An example of the case of LR fuzzy random sets is given.

Keywords: Fuzzy random sets, linear models, Bootstrap distribution.

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]. The random elements considered actually in many practical application 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. There are two main lines concerning regression modeling with fuzzy data in literature: namely, the so-called fuzzy or possibilistic regression proposed by Tanaka [11] and widely analyzed since then [4,5,11] and the so-called least squares problems of linear models [1,7,9,8,14] with fuzzy random sets [2,10,12,13]). In the former research line, the regression models are established based on possibilistic inclusion relationship between input and output of the systems rather than stochastic statistical settings. The last research line is based on statistical nonparametric settings, to consider both effects of randomness and fuzziness to the systems in the regression modeling, and the parameters (vector valued or fuzzy sets valued) estimation of the linear models are solved with least squares methods under metric between sets (see [1,2,7,9,8,14] and literature therein), and some concrete computational formulas for parameter estimation for simple linear regression model have been given. However, the same problems remain to be further investigated for the case of multivariate linear regression with fuzzy random sets.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_24

In this paper, we focus on a simple two-variate linear regression model with fuzzy random sets under concepts of functional data analysis. Based on the support function of the fuzzy random sets, we treat the fuzzy random sets as special functional data, and estimate the linear model within the support functional space. An example of the case of LR fuzzy random sets is given.

2 Preliminaries

2.1 Fuzzy Set on $\mathbb{R}^n$

A fuzzy set $\tilde{u}$ of $\mathbb{R}^n$ equivalents to its membership function $\tilde{u} : \mathbb{R}^n \to [0, 1]$, where the number $\tilde{u}(x)$ represents the degree of membership that x belongs to $\tilde{u}$. By $F(\mathbb{R}^n)$ we denote the collection of all normal, convex and compact fuzzy sets on $\mathbb{R}^n$, i.e. for $\tilde{u} \in F(\mathbb{R}^n)$, (1) There exists $x_0 \in \mathbb{R}^n$ such that $\tilde{u}(x_0) = 1$; (2) The $\alpha-$cut of $\tilde{u}$, $\tilde{u}_\alpha := \{x \in \mathbb{R}^n : \tilde{u}(x) \geq \alpha\}$, $\alpha \in (0, 1]$, is a convex and compact set of $\mathbb{R}^n$; (3) $\tilde{u}_0 := cl\{x \in \mathbb{R}^n : \tilde{u}(x) > 0\}$, the support of $\tilde{u}$, is compact.

Zadeh's extension principle [4] allows us to define addition and scalar multiplication on $F(\mathbb{R}^n)$:

$$(\tilde{u} \oplus \tilde{v})(x) = \sup_{s+t=x} \min(\tilde{u}(s), \tilde{v}(t)), x \in \mathbb{R}^n.$$

$$(a \odot \tilde{u})(x) = \begin{cases} \tilde{u}(\frac{x}{a}), a \neq 0 \\ 0, a = 0 \end{cases} \quad a \in \mathbb{R}.$$

and [9] for any $a, b \in \mathbb{R}$, it holds

$$(ab) \odot \tilde{u} = a \odot (b \odot \tilde{u}), a \odot (\tilde{u} \oplus \tilde{v}) = (a \odot \tilde{u}) \oplus (a \odot \tilde{v}).$$

But it holds only for $ab \geq 0, a, b \in \mathbb{R}$

$$(a + b) \odot \tilde{u} = (a \odot \tilde{u}) \oplus (b \odot \tilde{u}).$$

It indicates that $(F(\mathbb{R}^n), \oplus, \odot)$ is not a linear space. With Minkowski's sets operation it holds

$$(\tilde{u} \oplus \tilde{v})_\alpha = \tilde{u}_\alpha \oplus \tilde{v}_\alpha, \quad \alpha \in (0, 1].$$

$$(a \odot \tilde{u})_\alpha = a \odot \tilde{u}_\alpha, \quad \alpha \in (0, 1].$$

Definition 2.1 [14,2]. For $\tilde{u}, \tilde{v} \in F(\mathbb{R}^n)$, if there exists $\tilde{h} \in F(\mathbb{R}^n)$ such that $\tilde{u} = \tilde{v} \oplus \tilde{h}$, then $\tilde{h}$ is said to be Hukuhara difference between $\tilde{u}, \tilde{v}$ and denoted by $\tilde{h} := \tilde{u} \ominus_H \tilde{v}$.

The support function of $\tilde{u} \in F(\mathbb{R}^n)$ is defined as

$$S_{\tilde{u}_\alpha}(x) = \begin{cases} \sup_{t \in \tilde{u}_\alpha} \{x \cdot t\}, \alpha \in (0, 1], \\ 0, \alpha = 0. \end{cases} \quad x \in S^{n-1} = \{x : \| x \| = 1\}.$$

where $\cdot$ denotes the inner product in the Euclidean space $\mathbb{R}^n$. It holds that for $\tilde{u}, \tilde{v} \in F(\mathbb{R}^n)$ and $a \in \mathbb{R}$,

$$S_{\tilde{u} \oplus \tilde{v}} = S_{\tilde{u}} + S_{\tilde{v}}.$$

$$S_{a\odot\tilde{u}}(x) = aS_{\tilde{u}}(x), a > 0; S_{a\odot\tilde{u}}(x) = -aS_{\tilde{u}}(-x), a < 0.$$

thus, it holds that

$$S_{((a\odot\tilde{u})\oplus(b\odot\tilde{v}))_\alpha}(x) = \begin{cases} (aS_{\tilde{u}_\alpha} + bS_{\tilde{v}_\alpha})(x), a, b > 0 \\ -(aS_{\tilde{u}_\alpha} + bS_{\tilde{v}_\alpha})(-x), a, b < 0. \end{cases}$$

where $\alpha \in [0, 1]$. Thus, the map $S : F(\mathbb{R}^n) \to L^2(S^{n-1} \times [0, 1])$, $\tilde{u} \mapsto S_{\tilde{u}_\alpha}(x)$ enables us to view the fuzzy set $\tilde{u}$ as a support function equivalently, i.e. the map S embeds $F(\mathbb{R}^n)$ into a cone of functional Hilbert space [7].

We will employ the distance between $\tilde{u}, \tilde{v}$ proposed by [4] by the L_2 metric δ_2,

$$\delta_2(\tilde{u}, \tilde{v}) := \Big(n \int_0^1 \int_{S^{n-1}} (S_{\tilde{u}_\alpha}(x) - S_{\tilde{v}_\alpha}(x))^2 \mu(dx) d\alpha\Big)^{1/2},$$

where μ is a normalized Lebesgue measure. This distance has been widely used in area of fuzzy set-valued analysis, and in recent years several alternative versions of which as new metrics between fuzzy values have been proposed in literature, see [2,13].

2.2 Fuzzy Random Sets (Fuzzy Random Variables)

Fuzzy random sets as an extension of the concept of random sets had been introduced by Puri and Ralescu [10], and other definitions of fuzzy random sets were also proposed by Kwakernaak, Kruse and Meyer and Krätschmer [9] in different setting.

Definition 3.1 [10]. Let $(\Omega, \mathcal{B}, P)$ be a complete probability space. The mapping $\tilde{X} : \Omega \to F(\mathbb{R}^n)$ is said to be a fuzzy random set (frs) if $\tilde{X}$ is $\mathcal{B} - \mathcal{A}$ measurable, where we assume $\mathcal{A}$ is a σ-algebra induced by $\tilde{X}$ associated with δ_2.

Let $\tilde{X}$ be a frs, then for $\alpha \in [0, 1]$, $S_{\tilde{X}_\alpha}$ is a special random element, for a fixed $x \in S^{n-1}$, $S_{\tilde{X}_\alpha}(x)$ is random variable: $\Omega \to \mathbb{R}$, $\omega \mapsto S_{\tilde{X}_\alpha(\omega)}(x)$. A sample $\tilde{x}$ from $\tilde{X}$ can be viewed as a fuzzy data, thus, $S_{\tilde{x}}$ is a special functional data, an equivalence of $\tilde{x}$ [7].

Definition 3.2 [10]. Let $\tilde{X}$ be a frs. The Aumann expectation of $\tilde{X}$ is defined as a fuzzy set $E\tilde{X} \in F(\mathbb{R}^n)$ satisfying

$$\forall\alpha \in [0, 1] : (E\tilde{X})_\alpha = E(\tilde{X}_\alpha),$$

Here $E(\tilde{X}_\alpha)$ is the Aumann expectation of the random set $\tilde{X}_\alpha$ defined by

$$E(\tilde{X}_\alpha) = \{E\eta : \eta(\omega) \in \tilde{X}_\alpha(\omega)\ P - a.e.\ and\ \eta \in L^1(\Omega, \mathcal{B}, P)\}.$$

Note that $E(S_{\tilde{X}_\alpha}) = S_{E(\tilde{X}_\alpha)}$ [13,14] if the expectation $E(\tilde{X}_\alpha)$ exists, where $E(\tilde{X}_\alpha)$ is an Aumann expectation of $(\tilde{X}_\alpha), \alpha \in [0, 1]$ [10,9].

In the sequel, we assume that frs $\tilde{X}$ is with second order, i.e.

$$E(\|\tilde{X}\|) := E(\delta_2^2(\tilde{X}, \{0\})) < +\infty,$$

The Fréchet variance of $\tilde{X}$ w.r.t distance δ_2 is given in [12] as

$$Var(\tilde{X}) := E(\delta_2^2(\tilde{X}, E(\tilde{X}))) = n \int_0^1 \int_{S^{n-1}} Var(S_{\tilde{X}_\alpha}(x))\mu(dx)d\alpha.$$

and the Fréchet covariance of frs's $\tilde{X}, \tilde{Y}$ is also given in [12] as

$$Cov(\tilde{X}, \tilde{Y}) := n \int_0^1 \int_{S^{n-1}} Cov(S_{\tilde{X}_\alpha}(x), S_{\tilde{Y}_\alpha}(x))\mu(dx)d\alpha.$$

Note that,

$$Cov((a \odot \tilde{X}) \oplus (b \odot \tilde{Y}), c \odot \tilde{Z}) = acCov(\tilde{X}, \tilde{Z}) + bcCov(\tilde{Y}, \tilde{Z})$$

holds only for $ac \geq 0, bc \geq 0, a, b, c \in \mathbb{R}$.

The independence of frs's can be followed by the independence of the random elements which is already defined by [13]. If two frs's $\tilde{X}$ and $\tilde{Y}$ are independent, then $Cov(\tilde{X}, \tilde{Y}) = 0$. However, if $Cov(\tilde{X}, \tilde{Y}) \neq 0$, then $\tilde{X}$ and $\tilde{Y}$ will be dependent in some sense of semi-linear or non-linear [3].

Remark 2.1. The Fréchet variance, covariance can be defined w.r.t. different distances for frs (see [2,4,13]), and in general these distances such as $d_\infty, \delta_2, D_\theta^\varphi$ [2,4,13] are not coincide each other except some special cases. We prefer to employ Näther's one since that the concerned distance δ_2 is standard and simple one used in functional analysis.

Fréchet Principle [12]. The $E(\tilde{X})$ is the solution of the optimization problem $\inf_{\tilde{Y} \in F(\mathbb{R}^n)} E(\delta_2^2(\tilde{X}, \tilde{Y}))$.

Let $\tilde{X}, \tilde{Y}$ be frs's, and let $\{\tilde{X}_i\}, \{\tilde{Y}_i\}, i = 1, \cdots, m$, be independent observations on $\tilde{X}, \tilde{Y}$, respectively. Then equivalently we have r.v. $S_{\tilde{X}_\alpha}(x), S_{\tilde{Y}_\alpha}(x)$ and the functional data sets $\{S_{\tilde{X}_{i\alpha}}(x)\}$, $\{S_{\tilde{Y}_{i\alpha}}(x)\}$, $i = 1, \cdots, m$, and the estimations of $E(S_{\tilde{X}_\alpha}(x))$, $Var(S_{\tilde{X}_\alpha}(x))$ and $Cov(S_{\tilde{X}_\alpha}(x), S_{\tilde{Y}_\alpha}(x))$ are respectively as follows,

$$E(\widehat{S_{\tilde{X}_\alpha}(x)}) = \frac{1}{m}\sum_{i=1}^m S_{\tilde{X}_{i\alpha}}(x), Var\widehat{(S_{\tilde{X}_\alpha}}(x)) = \frac{1}{m}\sum_{i=1}^m (S_{\tilde{X}_{i\alpha}}(x) - S_{\overline{\tilde{X}}_\alpha})^2,$$

$$Cov(S_{\tilde{X}_\alpha}\widehat{(x), S}_{\tilde{Y}_\alpha}(x)) = \frac{1}{m}\sum_{i=1}^m (S_{\tilde{X}_{i\alpha}}(x) - S_{\overline{\tilde{X}}_\alpha})(S_{\tilde{Y}_{i\alpha}}(x) - S_{\overline{\tilde{Y}}_\alpha}).$$

So that $\widehat{E\tilde{X}} = n \int_0^1 \int_{S^{n-1}} E(\widehat{S_{\tilde{X}_\alpha}(x)})\mu(dx)d\alpha$, $\widehat{Var\tilde{X}} = n \int_0^1 \int_{S^{n-1}} Var\widehat{(S_{\tilde{X}_\alpha}}(x)) \cdot \mu(dx)d\alpha$, $\widehat{Cov(\tilde{X}, \tilde{Y})} = n \int_0^1 \int_{S^{n-1}} Cov(S_{\tilde{X}_\alpha}\widehat{(x), S}_{\tilde{Y}_\alpha}(x))\mu(dx)d\alpha$.

3 A Simple Multivariate Linear Regression Model with frs

Now we consider a new two-variate linear model with frs's, i.e. the case where the response frs $\tilde{Y}$ can be approximately linearly expressed by two explanatory frs's $\tilde{x}_1, \tilde{x}_2$ (compare with the considered models in [1,9,8,14]),

$$\tilde{Y} = \tilde{a} \oplus \beta_1 \tilde{x}_1 \oplus \beta_2 \tilde{x}_2 \oplus \tilde{\varepsilon}, \tag{1}$$

where $\tilde{a}$ is a fuzzy number to be estimated, β_1, β_2 are real number- valued parameters to be estimated, $\tilde{\varepsilon}$ is a uncertain disturbance frs with unknown probability distribution, whose Aumann expectation is assumed to be $E(\tilde{\varepsilon}) = \tilde{0}$, which means that given the realization $\tilde{x}_1^0, \tilde{x}_2^0$ of $\tilde{x}_1, \tilde{x}_2$

$$E(\tilde{Y}|\tilde{x}_1^0, \tilde{x}_2^0) = \tilde{a} \oplus \beta_1 \tilde{x}_1^0 \oplus \beta_2 \tilde{x}_2^0 \oplus \tilde{0}. \tag{2}$$

We assume that for the model there exists Hukuhara difference $\tilde{Y} \ominus_H (\tilde{a} \oplus \beta_1 \tilde{x}_1 \oplus \beta_2 \tilde{x}_2)$ and frs $\tilde{\varepsilon}$ can be formally expressed as

$$\tilde{\varepsilon} = \tilde{Y} \ominus_H (\tilde{a} \oplus \beta_1 \tilde{x}_1 \oplus \beta_2 \tilde{x}_2). \tag{3}$$

such that $\tilde{Y} = (\tilde{a} \oplus \beta_1 \tilde{x}_1 \oplus \beta_2 \tilde{x}_2) \oplus (\tilde{Y} \ominus_H (\tilde{a} \oplus \beta_1 \tilde{x}_1 \oplus \beta_2 \tilde{x}_2)))$.

Assume that we have independent observation $\{\tilde{Y}_i\}$,$\{\tilde{x}_{1i}\}$,$\{\tilde{x}_{2i}\}$ on $\tilde{Y}, \tilde{x}_1, \tilde{x}_2$, respectively, equivalently we have three functional data sets $\{S_{\tilde{Y}_{i\alpha}}(x)\}$, $\{S_{\tilde{x}_{1i\alpha}}(x)\}$, $\{S_{\tilde{x}_{2i\alpha}}(x)\}$, $i = 1, \cdots, m$.

Theorem 3.1. The least squares problem

$$\min_{\tilde{a} \in F(\mathbb{R}^n), \beta_1, \beta_2 \geqslant 0 \quad or \quad \beta_1, \beta_2 \leqslant 0} \frac{1}{m} \sum_{i=1}^{m} \delta_2^2(\tilde{Y}_i, \tilde{a} \oplus \beta_1 \tilde{x}_{1i} \oplus \beta_2 \tilde{x}_{2i})$$

has solutions (1) when $\beta_1, \beta_2 \geqslant 0$,

$$\hat{\beta}_1 = \max\Big\{0, \frac{\widehat{Cov(\tilde{Y}, \tilde{x}_1)} \widehat{Var \tilde{x}_2} - \widehat{Cov(\tilde{x}_1, \tilde{x}_2)} \widehat{Cov(\tilde{Y}, \tilde{x}_2)}}{\widehat{Var \tilde{x}_1} \widehat{Var \tilde{x}_2} - [\widehat{Cov(\tilde{x}_1, \tilde{x}_2)}]^2}\Big\},$$

$$\hat{\beta}_2 = \max\Big\{0, \frac{\widehat{Cov(\tilde{Y}, \tilde{x}_2)} \widehat{Var \tilde{x}_2} - \widehat{Cov(\tilde{x}_1, \tilde{x}_2)} \widehat{Cov(\tilde{Y}, \tilde{x}_1)}}{\widehat{Var \tilde{x}_1} \widehat{Var \tilde{x}_2} - [\widehat{Cov(\tilde{x}_1, \tilde{x}_2)}]^2}\Big\},$$

$$\hat{\tilde{a}} = \overline{\tilde{Y}} \ominus_H (\hat{\beta}_1 \overline{\tilde{x}}_1 \oplus \hat{\beta}_2 \overline{\tilde{x}}_2).$$

(2)when $\beta_1, \beta_2 \leqslant 0$,

$$\hat{\beta}_1 = \min\Big\{0, -\frac{\widehat{Cov(\tilde{Y}, -\tilde{x}_1)} \widehat{Var \tilde{x}_2} - \widehat{Cov(\tilde{x}_1, \tilde{x}_2)} \widehat{Cov(\tilde{Y}, -\tilde{x}_2)}}{\widehat{Var \tilde{x}_1} \widehat{Var \tilde{x}_2} - [\widehat{Cov(\tilde{x}_1, \tilde{x}_2)}]^2}\Big\},$$

$$\hat{\beta}_2 = \min\Big\{0, -\frac{\widehat{Cov(\tilde{Y}, -\tilde{x}_2)} \widehat{Var \tilde{x}_2} - \widehat{Cov(\tilde{x}_1, \tilde{x}_2)} \widehat{Cov(\tilde{Y}, -\tilde{x}_1)}}{\widehat{Var \tilde{x}_1} \widehat{Var \tilde{x}_2} - [\widehat{Cov(\tilde{x}_1, \tilde{x}_2)}]^2}\Big\},$$

$$\hat{\tilde{a}} = \overline{\tilde{Y}} \ominus_H (\hat{\beta}_1 (\overline{-\tilde{x}}_1) \oplus \hat{\beta}_2 (\overline{-\tilde{x}}_2)).$$

Proof. (1) Based on Fréchet principle, we have $\frac{1}{m} \sum_{i=1}^{m} \delta_2^2(\tilde{Y}_i, \tilde{a} \oplus \beta_1 \tilde{x}_{1i} \oplus \beta_2 \tilde{x}_{2i}) =$
$n \int_0^1 \int_{S^{n-1}} \frac{1}{m} \sum_{i=1}^{m} (S_{\tilde{Y}_{i\alpha}}(t) - S_{(\tilde{a} \oplus \beta_1 \tilde{x}_{1i} \oplus \beta_2 \tilde{x}_{2i})_\alpha}(t))^2 \mu(dt) d\alpha$
$=n \int_0^1 \int_{S^{n-1}} \frac{1}{m} \sum_{i=1}^{m} (S_{\tilde{Y}_{i\alpha}}(t) - S_{\tilde{a}_\alpha}(t) - \beta_1 S_{\tilde{x}_{1i\alpha}}(t) - \beta_2 S_{\tilde{x}_{2i\alpha}}(t))^2 \mu(dt) d\alpha$
$\geqslant n \int_0^1 \int_{S^{n-1}} \frac{1}{m} \sum_{i=1}^{m} (S_{\tilde{Y}_{i\alpha}}(t) - \beta_1 S_{\tilde{x}_{1i\alpha}}(t) - \beta_2 S_{\tilde{x}_{2i\alpha}}(t) - (S_{\overline{\tilde{Y}}_\alpha}(t) - \beta_1 S_{\overline{\tilde{X}}_{1\alpha}}(t) -$

$\beta_2 S_{\overline{\tilde{X}}_{2\alpha}}(t)))^2\mu(dt)d\alpha$, which means $\tilde{a} = \overline{\tilde{Y}} \ominus_H (\beta_1 \overline{\tilde{x}}_1 \oplus \beta_2 \overline{\tilde{x}}_2)$ minimizes $\frac{1}{m}\sum_{i=1}^{m} \delta_2^2 \cdot (\tilde{Y}_i, \tilde{a} \oplus \beta_1 \tilde{x}_{1i} \oplus \beta_2 \tilde{x}_{2i})$. Furthermore, set $f(\beta_1, \beta_2) := \frac{1}{m}\sum_{i=1}^{m} \delta_2^2(\tilde{Y}_i, (\overline{\tilde{Y}} \ominus_H (\beta_1 \overline{\tilde{x}}_1 \oplus \beta_2 \overline{\tilde{x}}_2)) \oplus (\beta_1 \tilde{x}_{1i} \oplus \beta_2 \tilde{x}_{2i})) = \widehat{Var\tilde{Y}} + \beta_1^2 \widehat{Var\tilde{x}_1} + \beta_2^2 \widehat{Var\tilde{x}_2} - 2\beta_1 \widehat{Cov(\tilde{x}_1, \tilde{Y})} - 2\beta_2 \widehat{Cov(\tilde{x}_2, \tilde{Y})} + 2\beta_1\beta_2 \widehat{Cov(\tilde{x}_1, \tilde{x}_2)}$, solving the equations $\frac{\partial f}{\partial \beta_1} = 0, \frac{\partial f}{\partial \beta_2} = 0$, then we have the solutions $\hat{\beta}_1, \hat{\beta}_2$ of (1).

The proof of (2) is analogous to the proof of (1), but we should take $\beta_1 \tilde{x}_1 = (-\beta_1)(-\tilde{x}_1), \beta_2 \tilde{x}_2 = (-\beta_2)(-\tilde{x}_1)$. □

4 Simulation Example

Assume that the observed human's pulse, diastolic pressure and systolic pressure can be comprehensively expressed by $\tilde{Y} = (\mu_y, l_y)_L$, $\tilde{x}_1 = (\mu_1, l_1)_L$, $\tilde{x}_2 = (\mu_2, l_2)_L$, the symmetric triangular fuzzy numbers (see.[13]), respectively, as shown in Table 1.

Table 1. Data of human's pulse, diastolic pressure and systolic pressure

i	$(\mu_y, l_y)_L$	$(\mu_1, l_1)_L$	$(\mu_2, l_2)_L$
1	(74,16)	(145.5, 27.5)	(85.5, 19.5)
2	(57.5, 10.5)	(132.5, 28.5)	(94.5, 23.5)
3	(73, 41)	(158.5, 27.5)	(85.5, 27.5)
4	(85.5, 24.5)	(131, 26)	(90, 28)
5	(75.5, 13.5)	(149.5, 29.5)	(76.5, 17.5)
6	(91, 28)	(147.5, 46.5)	(82, 34)
7	(73, 22)	(141.5, 32.5)	(89.5, 29.5)
8	(63.5, 14.5)	(169, 41)	(100.5, 24.5)
9	(55,12)	(119.5, 25.5)	(75.5, 28.5)
10	(78.5,23.5)	(174.5, 26.5)	(109, 21)
11	(65,13)	(165.5, 46.5)	(70, 23)
12	(69.5,14.5)	(150, 28)	(89, 16)
13	(81,20)	(158, 31)	(99.5, 25.5)
14	(78.5,13.5)	(163, 50)	(82, 30)
15	(52,14)	(173, 32)	(101, 32)
16	(60.5,12.5)	(134, 35)	(81, 28)
17	(78.5,19.5)	(158.5, 32.5)	(79, 19)
18	(73,14)	(150,51)	(88, 33)
19	(65.5,16.5)	(154.5, 66.5)	(65.5, 28.5)
20	(62.5,14.5)	(148,35)	(70, 15)

We obtain

$$\hat{\beta}_1 = 0.0236, \hat{\beta}_2 = 0.0865, \hat{\mu}_a = 59.6572, \hat{l}_a = 14.8489.$$

Then the concerned linear regression equation is

$$\hat{\tilde{Y}} = (59.6572, 14.8489)_L \oplus 0.0236\tilde{x}_1 \oplus 0.0865\tilde{x}_2.$$

However, for the obtained estimators of the model, the Hukuhara difference based residuals $\hat{\tilde{\varepsilon}} = (\mu_\varepsilon, l_\varepsilon)$ may not exist for some data. The residuals computed with the Hukuhara difference formula [13] are shown in Table 2, where some residuals (fuzzy data) with negative spreads appeared.

Table 2. Data of the residual $\hat{\tilde{\varepsilon}}$

i	$(\mu_\varepsilon, l_\varepsilon)$	i	$(\mu_\varepsilon, l_\varepsilon)$	i	$(\mu_\varepsilon, l_\varepsilon)$
1	(3.7707, -1.1851)	8	(-8.8413, -3.4364)	15	(-20.479, -4.3728)
2	(-13.4607, -7.0548)	9	(-14.01, -5.9166)	16	(-9.3281, -5.5976)
3	(2.2043, 23.1227)	10	(5.2934, 6.2087)	17	(8.2666, 2.2401)
4	(14.964, 6.6149)	11	(-4.6199, -4.9364)	18	(2.1886, -4.9078)
5	(5.6954, -3.5593)	12	(-1.3979, -2.3941)	19	(-3.471, -2.3844)
6	(20.7667, 9.1119)	13	(9.0048, 2.2131)	20	(-6.7069, -2.4729)
7	(2.2595, 3.8317)	14	(7.9009, -5.1247)		

Thus, there are only 7 values of the Hukuhara difference based residuals for the observations of Table 1, that is,

$$B = \{(2.2043, 23.1227), (14.964, 6.6149), (20.7667, 9.1119),$$
$$(2.2595, 3.8317), (5.2934, 6.2087), (9.0048, 2.2131), (8.2666, 2.2401)\}.$$

In the following we give an example of distributional simulation for the disturbance term $\tilde{\varepsilon}$. Taking B as a bootstrap population [6] and randomly resampling times of 10000. Using SAS on the bootstrap data, we output the histograms for center variable and spread variable respctively. The hypotheses about the distributions for center variable and spread variable remain to be tested in our future research.

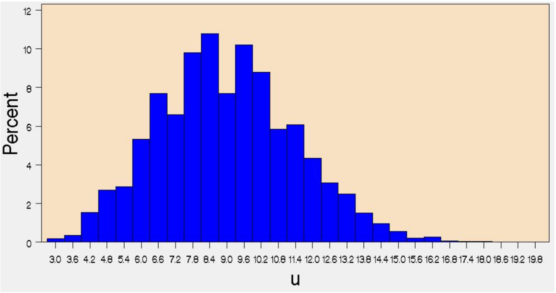


Fig. 1. The histogram for center variable of $\tilde{\varepsilon}$

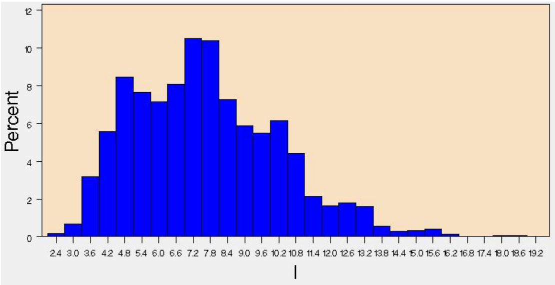


Fig. 2. The histograms histogram for spread variable of $\tilde{\varepsilon}$

Acknowledgment. The authors would like to thank the reviewers valuable comments. The research is supported by NSFC with grant number 11271096. Their financial support is gratefully acknowledged.

References

1. Sinova, B., Colubi, A., Gil, M.A., González-Rodríguez, G.: Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on choice of the metric. Information Sciences 199, 109–124 (2012)
2. Blanco-Fernández, A., Casals, M.R., Colubi, A., Corral, N., García-Bárzana, M., Gil, M.A., González-Rodríguez, G., López, M.T., De La Rosa De Sáa, S., Sinova, B.: Random fuzzy sets: A mathematical tool to develop statistical fuzzy data analysis. Iranian Journal of Fuzzy Systems 10(2), 1–28 (2013)
3. Cressie, N., Hulting, F.L.: A spatial statistical analysis of tumor growth. Journal of the American Statistical Association 87(418), 272–283 (1992)
4. Diamond, P., Kloeden, P.: Metric Spaces of Fuzzy Sets. World Scientific (1994)
5. Dimond, P., Körner, R.: Extended Fuzzy Linear Models and Least Squares Estimates. Computers and Mathematics with Applications 33, 15–32 (1997)
6. Efron and Tibshirani. An Introduction to the Bootstrap. Chapman & Hall (1993)
7. González-Rodríguez, G., Colubi, A., Gil, M.A.: Fuzzy data treated as functional data: A one-way ANOVA test approach. Computational Statistics and Data Analysis 56(4), 943–955 (2012)
8. Gil, M.A., Lubiano, M.A., Montenegro, M., López, M.T.: Least squares fitting of an affine function and strength of association for interval-valued data. Metrika 56, 97–111 (2002)
9. Krätschmer, V.: Least-squares estimation in linear regression models with vague concepts. Fuzzy Sets and Systems 157, 2579–2592 (2006)
10. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
11. Tanaka, H., Uejima, S., Asai, K.: Linear regression analysis with fuzzy model. IEEE Trans. Systems Man Cybernet. 12, 903–907 (1982)
12. Näther, W.: Regression with fuzzy random data. Computational Statistics & Data Analysis 51, 235–252 (2006)
13. Näther, W.: Random fuzzy variables of second order and applications to statistical inference. Information Sciences 133, 69–88 (2001)
14. Wünsche, A., Näther, W.: Least -squares fuzzy regression with fuzzy random variables. Fuzzy Sets and Systems 130, 43–50 (2002)

Cluster Analysis of Time Series via Kendall Distribution

Fabrizio Durante* and Roberta Pappadà

School of Economics and Management,
Free University of Bozen–Bolzano, Bolzano, Italy
{fabrizio.durante,roberta.pappada}@unibz.it

Abstract. We present a method to cluster time series according to the calculation of the pairwise Kendall distribution function between them. A case study with environmental data illustrates the introduced methodology.

Keywords: Cluster analysis, Copula, Kendall distribution, Tail dependence.

1 Introduction

Cluster analysis plays an important role in extracting information from a group of different time series. It can be used, for instance, to find some dependence information, which is a key tool in geosciences and hydrology in order to understand the relationships between different variables. In general, a time series clustering procedure involves the choice of an adequate metric between the univariate time series, which allows to group together series exhibiting common trends occurring at different times or similar sub-patterns in the data, according to the idea of similarity one has adopted (see [1]).

A widely used approach to measure similarity is to consider a Pearson-correlation based distance metric. However, recent studies have underlined that classical correlation measures are often inadequate to capture the real dependence structure between individual risk factors, especially in a financial and environmental context (see, for instance, [2], [3]). As such, several investigations have been carried out during the last years from different perspectives, exploiting tools from extreme-value analysis ([4], [5]) to the concept of tail copulas (see, for instance, [6] and the references therein). In particular, many research efforts have remarked on the usefulness of extreme value theory in assessing climate changes and detecting spatial clusters (see, for instance [7], [8]). Moreover, recent developments in statistical hydrology have shown the great potential of copulas for the construction of multivariate cumulative distribution functions and for carrying out a multivariate frequency analysis ([9], [10]). Extreme value copulas have been largely used to investigate the spatial dependencies between

* This work was supported by Free University of Bozen-Bolzano via the project MODEX.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_25

the involved variables, introducing a novel contribution to the interpretation of meteorological and hydrological phenomena ([11], [12], [13]). From another perspective, methods have been recently proposed in order to cluster time series observations according to a suitable copula-based dissimilarity measure, with applications in the financial setting. Such an approach has been adopted, for instance, in [14] focusing on the use of conditional Spearman's correlation, and in [15], [16] where the clustering procedure is based on the estimation of pairwise tail dependence coefficients.

Management of environmental resources often requires the analysis of spatial rainfall extremes which typically exhibit some form of dependence as a result of the regional nature of hydrological phenomena. Reliable estimates of extreme rainfall events are required for several hydrological purposes and their spatial distribution is of both physical and practical interest, particularly in the case of regional studies. Several approaches are available in the literature for the characterization of spatial extremes, relying on a likelihood-based approach ([17], [18]), a Bayesian approach ([19]) and cluster analysis for assessing the spatial distribution of extremes ([20], [21]). In particular, the detection of spatial clusters can help in summarizing available data, extracting useful information and formulating hypothesis for further research. Clustering could be used in order to identify homogeneous regions to be considered for regionalization procedures.

In the present contribution, we would like to use the Kendall distribution function associated with a random vector in order to develop a novel clustering procedure for grouping random vectors. We outline here briefly the possible application of the proposed methodology to hydrological data by analysing time series of maximum annual rainfall data collected at rain gauges of different sites in the province of Bolzano-Bozen (Italy). Notice that according to the approach in [22], homogeneity in the sense of Kendall's distance implies homogeneity in the sense of return period, a notion frequently used in environmental sciences for the identification of dangerous events and risk assessment (see also [23],[24]).

2 Clustering via Kendall Distribution

We recall that a (bivariate) copula is a joint cumulative probability distribution function with uniform univariate margins on $\mathbb{I} = [0, 1]$. If we consider a random pair (X, Y) with cumulative continuous distribution function H, then the bivariate probability integral transform is the random variable defined by $W = H(X, Y)$. It is known that W just depends on the copula C of (X, Y) and it is equal in distribution to $C(U, V)$, where $U = F_X(X)$ and $V = F_Y(Y)$, being F_X, F_Y the univariate marginals of X and Y, respectively. First introduced in [25] for inference on Archimedean copulas, the Kendall distribution function (see also [26]) is simply the distribution function of W and is given by

$$K(q) = \mathbb{P}(W \leq q),$$

where $q \in [0, 1]$ is a probability level.

There are two important particular cases for the Kendall distribution. When X and Y are comonotonic, one finds $K(q) = K_M(q) = q$ for all $0 \leq q \leq 1$, which corresponds to $C(u, v) = M(u, v) = \min(u, v)$, where M is the the Fréchet-Hoeffding upper bound copula. Under the hypothesis of independence between X and Y, which is equivalent to consider $C(u, v) = \Pi(u, v) = uv$, K has the form $K(q) = K_\Pi(q) = q - q\log(q)$, $0 \leq q \leq 1$. Thus, on the graph of K based on pseudo-random samples from a positively dependent bivariate vector (X, Y), perfect positive dependence would translate into data points aligned on the line $y = x$, while the plot will be seen to match nearly the curve $K_\Pi(q)$ as the data become less and less dependent. Notice that, for each Kendall distribution K, one has the lower bound $K \geq K_M$ on $\mathbb{I}$. Starting with [27] (see also [28]), ordering properties of Kendall distributions have been used to detect dependence in copula models. Here we show how to use them to provide a clustering procedure for time series.

Suppose that we have at disposal a set of time series $X_1^t, \ldots, X_n^t$, corresponding to n different measurements collected at time $t \in \{1, \ldots, T\}$. Such time series are assumed to be a random sample from an unknown vector $\mathbf{X} = (X_1, \ldots, X_n)$. In order to interpret properly the following results it is also convenient to suppose that the all the pairs in $\mathbf{X}$ are positively quadrant dependent, i.e. their copula is grater than or equal to Π. We would like to group the components of $\mathbf{X}$ according to the strength of their inter–dependence. To do this, following the general principle applied in [14], we may proceed as follows:

1. Calculate the Kendall distribution function $K(\cdot)$ for each pair (X_i, X_j), and denote it by K^{ij}.
2. Define a kind of distance between X_i and X_j in terms of the related Kendall distribution $K^{ij} = K$ and the Kendall distribution K_M of comonotone random variables by one of the following definitions:

$$d_2(K, K_M) = \int_0^1 (q - K(q))^2 dq$$
$$d_\infty(K, K_M) = \sup_{q\in[0,1]} |q - K(q)| dq$$

   Intuitively, two random variables have small distance if their Kendall distribution is close to K_M or, in other words, if they tend to be comonotone.
3. From these metrics, create a suitable dissimilarity matrix $D := (\delta_{ij})$, $i, j = 1, \ldots, n$, for instance by using $\delta_{ij} = d_2(K^{ij}, K_M)$. In fact, if the random variables are comonotone, their dissimilarity is 0, while this number increases when they are becoming less and less dependent. Hence, in this construction, the larger the distance, the weaker the dependence.
4. Apply classical cluster techniques to the obtained dissimilarity matrix. In particular, agglomerative hierarchical methods with nearest distance (single linkage), furthest distance (complete linkage) and average distance (average linkage) can be used as grouping criteria.

For what concerns the estimation procedure of the Kendall distribution function we rely on non-parametric estimation by using the empirical distribution function computed as in [29]. Suppose that $(X_{11}, X_{12}), \dots, (X_{T1}, X_{T2})$ is a random sample from a distribution H with copula C. The empirical Kendall distribution function K_T is given, for all $q \in [0,1]$, by

$$K_T(q) = \frac{1}{T}\sum_{j=1}^{T} \mathbf{1}(W_j \le q),$$

where, for each $j \in \{1, \dots, T\}$,

$$W_j = \frac{1}{T+1}\sum_{t=1}^{T} \mathbf{1}(X_{t1} < X_{j1}, X_{t2} < X_{j2}).$$

The limiting behaviour of the empirical process $\sqrt{T}(K_T - K)$ has been discussed in [30], where the convergence in law to a centered Gaussian limit under mild regularity conditions is proved.

3 An Empirical Case Study

In order to briefly illustrate a possible application of the proposed methodology we present here a case study from environmental data. The data were collected by "Ufficio Idrografico" of the province of Bolzano-Bozen and are available online. They are related to daily rainfall measurements recorded at 18 gauge stations spread across the province of Bolzano-Bozen in the North-Eastern Italy. This results in a set of $d = 18$ time series originally formed by $T = 18262$ observations. Tab. 1 reports the available information on the analysed rainfall records. From these time series, we extracted annual maxima at each spatial location resulting in a 50×18 matrix of time series observations $\tilde{X}_1^m, \dots, \tilde{X}_d^m$, $m \in \{1, \dots, 50\}$, summarized by Fig. 1. The selection of annual maxima has two main goals: it transforms data with strong seasonality into data that can be assumed to be independent and identically distributed; it transforms data that may have a general dependence structure into data that are positively dependent (actually, they are coupled by an extreme-value copula). For more details, see [5]. The latter property is quite relevant since it allows to apply the method described in Section 2 in order to detect the presence of clusters of the analysed sites on the basis of the componentwise maxima.

Specifically, we compute the dissimilarity matrix $D := (\delta_{ij})$, $i, j = 1, \dots, d$, such that the dissimilarity between two time series is defined as the distance

$$\delta_{ij} = d_2(\hat{K}^{ij}, K_M) = \int_0^1 (q - \hat{K}^{ij}(q))^2 dq,$$

where $\hat{K}^{ij}$ is the empirical Kendall distribution function based on the maxima observations $(\tilde{X}_i^m, \tilde{X}_j^m)$, $m \in \{1, \dots, 50\}$.

The choice of this metric reflects the final goal of the clustering procedure in the sense that two strongly dependent time series will give an extremely low

Table 1. Summary of the rainfall measurement stations

Code	Station	Longitude	Latitude	Height (m)
0220	S.VALENTINO ALLA MUTA	10.5277	46.7745	1520
0310	TUBRE	10.4775	46.6503	1119
2090	PLATA	11.1783	46.8225	1147
3140	FLERES	11.3477	46.9639	1246
3260	VIPITENO-CONVENTO	11.4295	46.8978	948
8320	BOLZANO	11.3127	46.4976	254
9150	SESTO	12.3477	46.7035	1310
0250	MONTE MARIA	10.5213	46.7057	1310
0480	MAZIA	10.6175	46.6943	1570
1580	VERNAGO	10.8493	46.7357	1700
2170	S.LEONARDO PASSIARIA	11.2471	46.8091	644
2670	PAVICOLO	11.1093	46.6278	1400
3450	RIDANNA	11.3068	46.9091	1350
4450	S.MADDALENA IN CASIES	12.2427	46.8353	1398
6650	FUNDRES	11.7029	46.8872	1159
8570	BRONZOLO	11.3111	46.4065	226
8730	REDAGNO	11.3968	46.3465	1562
9100	ANTERIVO	11.3678	46.2773	1209

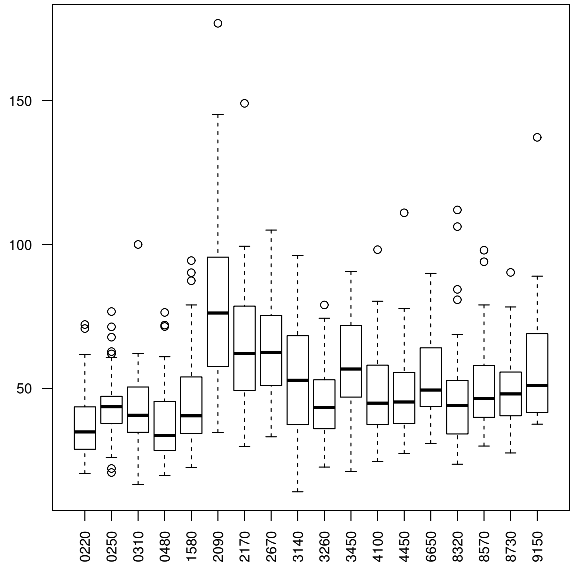


Fig. 1. Boxplot of annual maxima at each station from 1961 to 2010. The station codes are as Tab. 1. On the y-axis the amount of rainfall is measured in millimeters.

value of their dissimilarity. The results of the clustering procedure are illustrated by a tree diagram usually referred to as dendrogram, which represents the arrangement of the clusters produced by hierarchical agglomerative clustering. In Fig. 2, the dendrogram based on complete linkage is displayed. The vertical axis represents the distance at which two clusters are joined. From the dendrogram it is possible to identify, e.g., four different groups, by cutting at about height 0.06.

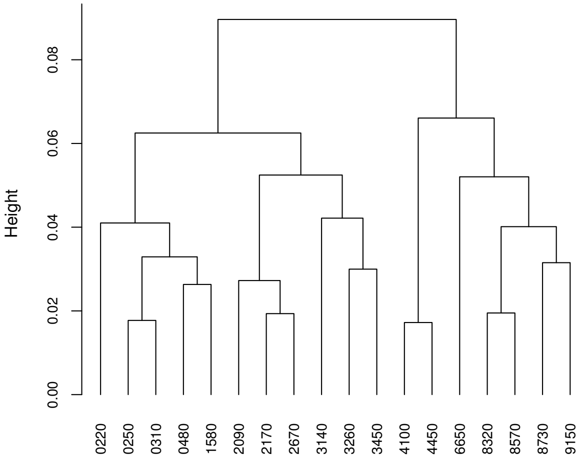


Fig. 2. Dendrogram for the 18 rainfall measurement stations listed in Tab. 1 based on the complete linkage method

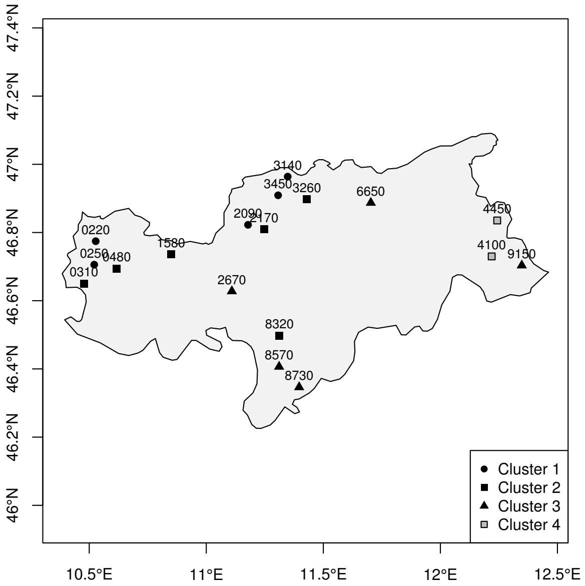


Fig. 3. Map of the rainfall measurement stations marked according the the 4-clusters solution in the province of Bolzano–Bozen (North-Eastern, Italy)

The 4-clusters solution is visualized on the map in Fig. 3, where the stations are marked according to their cluster.

For the hydrological interpretation of the results it seems that several factors should be taken into account in order to determine correlated rainfall extremes.

In fact, not only the geographical proximity plays a role, but also the strong heterogeneity in morphological and climatic features.

4 Conclusions

We have presented a procedure for grouping time series according to a copula-based dependence function among them. In particular, we considered a dissimilarity measure that is based on the Kendall distribution associated to two continuous random variables, since such a function provides useful information in terms of environmental risk, as shown in [22]. The proposed approach complements similar methods provided by the authors about copula-based clustering of time series (see, e.g., [14], [16]).

References

1. Liao, T.W.: Clustering of time series data - a survey. Pattern Recogn. 38(11), 1857–1874 (2005)
2. Embrechts, P., McNeil, A.J., Straumann, D.: Correlation and Dependence in Risk Management: Properties and Pitfalls. Cambridge Univ. Press, New York (2001)
3. Poulin, A., Huard, D., Favre, A.-C., Pugin, S.: Importance of tail dependence in bivariate frequency analysis. J. Hydrol. Eng. 12, 394–403 (2007)
4. Gudendorf, G., Segers, J.: Extreme-value copulas. In: Jaworski, P., Durante, F., Härdle, W., Rychlik, T. (eds.) Copula Theory and its Applications. Lecture Notes in Statistics - Proceedings, vol. 198, pp. 127–145. Springer, Heidelberg (2010)
5. Salvadori, G., De Michele, C., Kottegoda, N.T., Rosso, R.: Extremes in Nature. An Approach Using Copulas. Water Sci. and Technology Library 56. Springer (2007)
6. Jaworski, P.: Tail behaviour of copulas. In: Jaworski, P., Durante, F., Härdle, W., Rychlik, T. (eds.) Copula Theory and its Applications. Lecture Notes in Statistics - Proceedings, vol. 198, pp. 161–186. Springer, Heidelberg (2010)
7. Gaetan, C., Grigoletto, M.: A hierarchical model for the analysis of spatial rainfall extremes. J. ABES 12(4), 434–449 (2007)
8. Scotto, M.G., Barbosa, S.M., Alonso, A.M.: Extreme value and cluster analysis of European daily temperature series. Journal of Applied Statistics 38(12), 2793–2804 (2011)
9. Favre, A.-C., Adlouni, S.E., Perreault, L., Thiemonge, N., Bobee, B.: Multivariate hydrological frequency analysis using copulas. Water Resour. Res. 40 (2004)
10. Salvadori, G., De Michele, C.: Frequency analysis via copulas: theoretical aspects and applications to hydrological events. Water Resour. Res. 40 (2004)
11. Bárdossy, A.: Copula-based geostatistical models for groundwater quality parameters. Water Resour. Res. 42(11) (2006)
12. Bonazzi, A., Cusack, S., Mitas, C., Jewson, S.: The spatial structure of European wind storms as characterized by bivariate extreme-value Copulas. Nat. Hazards Earth Syst. Sci. 12, 1769–1782 (2012)
13. Genest, C., Favre, A.-C.: Everything you always wanted to know about copula modeling but were afraid to ask. J. Hydrologic Eng. 12(4), 347–368 (2007)
14. Durante, F., Pappadà, R., Torelli, N.: Clustering of financial time series in risky scenarios. Adv. Data Anal. Classif. (2013) (in press), doi: 10.1007/s11634-013-0160-4

15. De Luca, G., Zuccolotto, P.: A tail dependence-based dissimilarity measure for financial time series clustering. Adv. Data Anal. Classif. 5(4), 323–340 (2011)
16. Durante, F., Pappadà, R., Torelli, N.: Clustering of extreme observations via tail dependence estimation. Statist. Papers (in press, 2014)
17. Buishand, T., de Haan, L., Zhou, C.: On spatial extremes: With application to a rainfall problem. Ann. Appl. Statist. 2, 624–642 (2008)
18. Cooley, D., Naveau, P., Poncet, P.: Variograms for spatial max-stable random fields. In: Dependence in Probability and Statistics. Lectures Notes in Statistics, pp. 373–390. Springer, Heidelberg (2006)
19. Cooley, D., Nychka, D., Naveau, P.: Bayesian spatial modeling of extreme precipitation return levels. J. Amer. Statist. Assoc. 102, 824–840 (2007)
20. Robeson, S.M., Doty, J.A.: Identifying rogue air temperature stations using cluster analysis of percentile trends. J. Climate 18, 1275–1287 (2005)
21. Scotto, M.G., Alonso, A.M., Barbosa, S.M.: Clustering time series of sea levels: Extreme value approach. J. Waterway, Port, Coastal, and Ocean Engrg. 136, 215–225 (2010)
22. Salvadori, G., De Michele, C., Durante, F.: On the return period and design in a multivariate framework. Hydrol. Earth Syst. Sci. 15, 3293–3305 (2011)
23. Salvadori, G., Durante, F., De Michele, C.: Multivariate return period calculation via survival functions. Water Resour. Res. 49(4), 2308–2311 (2013)
24. Salvadori, G., Durante, F., Perrone, E.: Semi–parametric approximation of the Kendall's distribution and multivariate return periods. J. SFdS 154(1), 151–173 (2013)
25. Genest, C., Rivest, L.-P.: Statistical inference procedures for bivariate Archimedean copulas. J. Amer. Statist. Assoc. 88(423), 1034–1043 (1993)
26. Genest, C., Rivest, L.-P.: On the multivariate probability integral transformation. Statist. Probab. Lett. 53(4), 391–399 (2001)
27. Capéraà, P., Fougères, A.-L., Genest, C.: A stochastic ordering based on a decomposition of Kendall's tau. In: Beneš, V., Štěpán, J (Eds.) Distributions with Given Marginals and Moment Problems. Kluwer Academic Publishers, Dordrecht, pp. 81–86
28. Nelsen, R.B., Quesada–Molina, J.J., Rodríguez–Lallena, J.A., Úbeda–Flores, M.: Kendall distribution functions. Statist. Probab. Lett. 65, 263–268 (2003)
29. Genest, C., Nešlehová, G., Ziegel, J., Inference, J.: in multivariate Archimedean copula models. TEST 20, 223–256 (2011)
30. Barbe, P., Genest, C., Ghoudi, K., Rémillard, B.: On Kendall's process. J. Multivar. Anal. 58(1996), 197–229 (1996)

Connectedness Measures of Spatial Contagion in the Banking and Insurance Sector

Fabrizio Durante[1,⋆], Enrico Foscolo[2], Piotr Jaworski[3], and Hao Wang[4]

[1] School of Economics and Management,
Free University of Bozen-Bolzano, Bolzano, Italy
fabrizio.durante@unibz.it
[2] School of Economics and Management,
Free University of Bozen-Bolzano, Bolzano, Italy
enrico.foscolo@unibz.it
[3] Institute of Mathematics,
University of Warsaw, Warszawa, Poland
p.jaworski@mimuw.edu.pl
[4] Department "Methods and Models for Economics, Territory and Finance",
Sapienza University of Rome, Rome, Italy
hao.wang@uniroma1.it

Abstract. We present some connectedness measures for an economic system that are derived from the spatial contagion measure. These measures are calculated directly from time series data and do not require any parametric assumption. The given definitions are illustrated in an empirical analysis of the behavior of European banking and insurance sector in the recent years.

Keywords: Contagion, Copula, Tail dependence.

1 Introduction

The recent financial crisis has renewed the interest in the interconnectedness among different financial institutions located in various countries. In particular, "systemic risk" has become a standard concept that relates to the risk imposed by interlinkages and interdependencies in a system or market, where the failure of a single entity or group of entities can cause potential difficulties to other entities, which could potentially bankrupt or bring down the entire system.

As stressed by [2], studies about systemic risk can be divided into two major groups. One approach consists of using network analysis and works directly on the structure and the nature of relationships between financial institutions in the market. Another approach investigates the impact of one institution on the market and its contribution to the global system risk [1]. Hence, the latter methodology requires the knowledge of the joint behavior of the financial institutions and is related to previous works about the so-called financial contagion

⋆ This work was supported by Free University of Bozen-Bolzano via the project MODEX.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_26

[9]. Roughly speaking, contagion refers to a significant increase in comovements of prices and quantities across markets, conditional on a crisis occurring in one market or group of markets.

Recently, the notion of financial contagion has been reformulated in terms of copulas in [7] (see also [4,6,8]). Specifically, it refers to the change of strength of dependence in the tail and in the center of the joint distribution associated with two financial positions. This concept has been further developed in [5], where a spatial contagion measure has been defined in order to quantify (in a normalized scale) the influence of one market over the others.

In this contribution, we review the notion of spatial contagion measure by pointing some of its main features. Then, inspired by [3], we define some simple measures of connectedness that can be used in order to investigate systemic risk in a set of financial institutions. The second part of the contribution is devoted to the empirical investigation of spatial contagion in a set of asset returns related to banking and insurance sectors, which have become increasingly interconnected especially during the last decade.

2 The Spatial Contagion Measure

The notion of spatial contagion measure has been introduced in [5]. Basically, it focuses on the discrepancies between tail and central sets of probability distribution function of two financial returns. This approach is based on the geometry of the underlying distribution and, for this reason, it is called spatial contagion. Formally, it is defined in the following way.

Let X and Y be two random variables on a suitable probability space representing the returns (or log-returns) of financial markets whose dependence is described by means of a copula C. Consider the following Borel sets of $\mathbb{R}^2$:

- the *tail set* T_{α_1,α_2} given by

$$T_{\alpha_1,\alpha_2} = [-\infty, q_X(\alpha_1)] \times [-\infty, q_Y(\alpha_2)],$$

  where $\alpha_1, \alpha_2 \in [0,1]$ and q_X and q_Y are the quantile functions associated with X and Y, respectively.
- the *central set* (or *mediocre set*) M_{β_1,β_2} given by

$$M_{\beta_1,\beta_2} = [q_X(\beta_1), q_X(1-\beta_1)] \times [q_Y(\beta_2), q_Y(1-\beta_2)]$$

  where $\beta_1, \beta_2 \in [0, 1/2]$.

Intuitively, T_{α_1,α_2} represents the "risky scenario" for the pair (X, Y), since it includes the bivariate observations that are less than a given threshold; while M_{β_1,β_2} represents the so-called "untroubled scenario", since it is related to all the observations that are in the central region of the joint distribution (being the extreme values excluded).

Definition 1. *Let $L \subseteq (0, 0.5)$. The (spatial) contagion measure from X to Y is defined by the formula*

$$\gamma(X \to Y) = \frac{1}{\lambda(L)} \lambda(\{\alpha \in L \mid \rho(T_{\alpha,1}) - \rho(M_{\alpha,0}) > 0\}), \tag{1}$$

where λ is the Lebesgue measure, $\rho(T_{\alpha,1})$ (respectively, $\rho(M_{\alpha,0})$) denotes the Spearman's correlation of the conditional distribution function of $[(X,Y) \mid (X,Y) \in T_{\alpha,1}]$ (respectively, $[(X,Y) \mid (X,Y) \in M_{\alpha,0}]$).

The introduced measure depends hence on the Spearman's rank correlation and avoids to restrict to the use of linear correlation. Moreover, the estimation of the contagion measure can be done mainly in a non-parametric way, as illustrated in [5].

Roughly speaking, the contagion measure counts how many times the correlation in the tail of the joint distribution is larger than the correlation in the central region for some predefined set of possible levels $\alpha \in L$, where L is usually chosen by the decision maker according to her/his risky attitude. Notice that the mapping $\alpha \to \varDelta_\alpha = \rho(T_{\alpha,1}) - \rho(M_{\alpha,0})$ for every $\alpha \in L = [a,b]$ depends only on the copula of (X,Y) (since it is based on rank correlation). Moreover, it can be positive, negative or changing in sign depending on the involved dependence. For instance:

- If (X,Y) has copula equal to the ordinal sum of comonotonicity and independence copula with respect to the partition $([0,a],[a,1])$, then $\varDelta_\alpha \geq 0$ for every $\alpha \in L$.
- If (X,Y) has copula equal to the ordinal sum of independence and comonotonicity copula with respect to the partition $([0,a],[a,1])$, then $\varDelta_\alpha \leq 0$ for every $\alpha \in L$.
- If (X,Y) has Gaussian copula with correlation $\rho > 0$, then $\varDelta_\alpha$ changes sign in $(0, 0.5)$ (see [8] for more details).

Starting with this definition, we consider now some derived measures of connectedness, inspired by the motivations presented in [3].

Let us consider historical time series from different asset returns $X_1, \ldots, X_d$ that are operating in the same sector and/or geographic region. Let $\mathcal{J}$ be a subset of indices in $\{1, 2, \ldots, d\}$. Let $\mu(X_i \to X_j)$ be the contagion measure from asset i to asset j. Then we can define the following measure of connectedness that may help in the identification of contagion effects from one asset to a group of assets or between groups of assets.

Definition 2. *Let $\mathcal{J}$ be a subset of indices in $\{1, 2, \ldots, d\}$, let $i \in \{1, \ldots, d\} \setminus \mathcal{J}$. We define contagion from X_i to $\{X_j : j \in \mathcal{J}\}$ as*

$$\mu(X_i \to \{X_j : j \in \mathcal{J}\}) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \mu(X_i \to X_j).$$

Obviously, $\mu(X_i \to \{X_j : j \in \mathcal{J}\}) = 0$ when $\mu(X_i \to X_j) = 0$ for each choice of the indices j's in $\mathcal{J}$. Analogously we can define the following measure.

Definition 3. *Let $\mathcal{I}, \mathcal{J}$ be disjoint subset of $\{1, \dots, d\}$. We define contagion from $\{X_i : i \in \mathcal{I}\}$ over $\{X_j : j \in \mathcal{J}\}$ as*

$$\mu(\{X_i : i \in \mathcal{I}\} \to \{X_j : j \in \mathcal{J}\}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mu(X_i \to \{X_j : j \in \mathcal{J}\}).$$

Both these measures of connectedness are obtained by aggregating spatial contagion measures at individual levels. They will be used below to provide some insights in understanding systemic risk.

3 Empirical Analysis

We consider the daily log-returns of the European banks and insurance companies characterizing the STOXX Europe 600 Index. The financial list counts 84 institutions and covers 16 countries[1] and 7 currencies[2]. The companies are divided into two sectors, Bank and Insurance, containing 47 and 37 assets, respectively. Moreover, each sector is divided into three groups according to the market capitalization. As result, we cluster the 84 assets by their capital size and sector into the following groups: 25 large banks, 8 medium banks, 14 small banks, 15 large insurance companies, 8 medium insurance companies, and 14 small insurance companies. Following [3], the emphasis on market returns is motivated by the desire to incorporate the most current information in our measures. On the other hand, the clustering procedure by market capitalization is designed for taking into account possible different trading liquidity and financial instability within groups of large-, medium- and small-sized companies.

The dataset refers to time interval January 3rd, 2005-December 31st, 2012. In order to compare the time-variation of the connectedness measures between the sectors and within the sectors, we fix two four-years time windows: 2005-2008 and 2009-2012; shortly, the "before the crisis" and the "after the crisis" period, respectively.

As mentioned in Section 2, the definitions of connectedness from one single financial institution to a sector of institutions and from one sector to another

Table 1. Contagion measure between different financial sectors in both periods using Definition 3. Letter B stands for the bank sector, while letter I for the insurance sector.

	$\mu(B \to B)$	$\mu(B \to I)$	$\mu(I \to B)$	$\mu(I \to I)$
2005-2008	0.43	0.37	0.40	0.32
2009-2013	0.23	0.24	0.20	0.23

[1] Shortly, AT, BE, CH, CZ, DE, DK, ES, FI, FR, GB, IE, IT, NL, NO, PT, and SE.

[2] British Pound, Czech Koruna, Danish Krone, Euro, Norwegian Krone, Swedish Krona, and Swiss Franc.

sector are based on the spatial contagion measure proposed in [5], which requires some ad-hoc algorithm to be computed. In this work, we calculate our measures directly on the data (without any preliminary ARMA-GARCH filter); for the computation, we refer to the Algorithm 4.2 and the Algorithm 4.3 as described in [5].

Firstly, for every single financial institution X_i, we calculate $\mu(X_i \to S)$ in the two considered periods, where S is formed by all institutions beloning to a specific sector (i.e., bank or insurance). In Figure 1, each line corresponds to the (smoothed) empirical density of the histograms related to all the measurements of type $\mu(X_i \to S)$ in a specific time period, where X_i is varying in a specific sector, while S equals bank, insurance, or both sectors. As can be seen, regardless of the choice of a different set S, the distribution of the spatial contagion measure moves towards smaller values during the second period. In order to highlight such a finding, we compute the contagion from one sector to another one in both periods; see Table 1. We note that, in the latter period, the overall contagion risk between financial institution sectors seems to be reduced. Nevertheless, for banks this change is more clear, since the two sets of three density curves seem to be unimodal; see the upper panel in Figure 1. For insurance companies, however, the evidence is different because after crisis the density curves seems to be bimodal, implying that for a large subset of these corporations contagion risk remains high; see the lower panel in Figure 1.

In order to give a graphical representation of the evolving relations, we provide a network diagram to show the linkages among different financial institutions by plotting a line connection when the contagion measure from institution X_i to institution X_j is larger that 0.5; see Figure 2. The charts highlight the fact that the contagion measures within and between sectors decreased after the crisis since the number of extreme edges in the previous period is much larger than in later period.

As can be noticed in Figure 2, when we look at the contagion within sectors before the crisis period, the large-sized banks are heavily connected, but they are less affected by medium- and small-sized banks. When we focus on the insurance sector, the situation is just the opposite. Here, large-sized insurance companies are more easily affected by medium- and small-sized companies, especially the small-sized ones. If we consider the contagion measure between sectors, contagion effects from banks to insurance companies are less likely than those from insurance companies to banks.

In the second period, however, all these effects seems to considerably reduce. A flight-to-quality evidence towards different markets and investments (e.g., bond market and cash equivalent) appears a possible explanation.

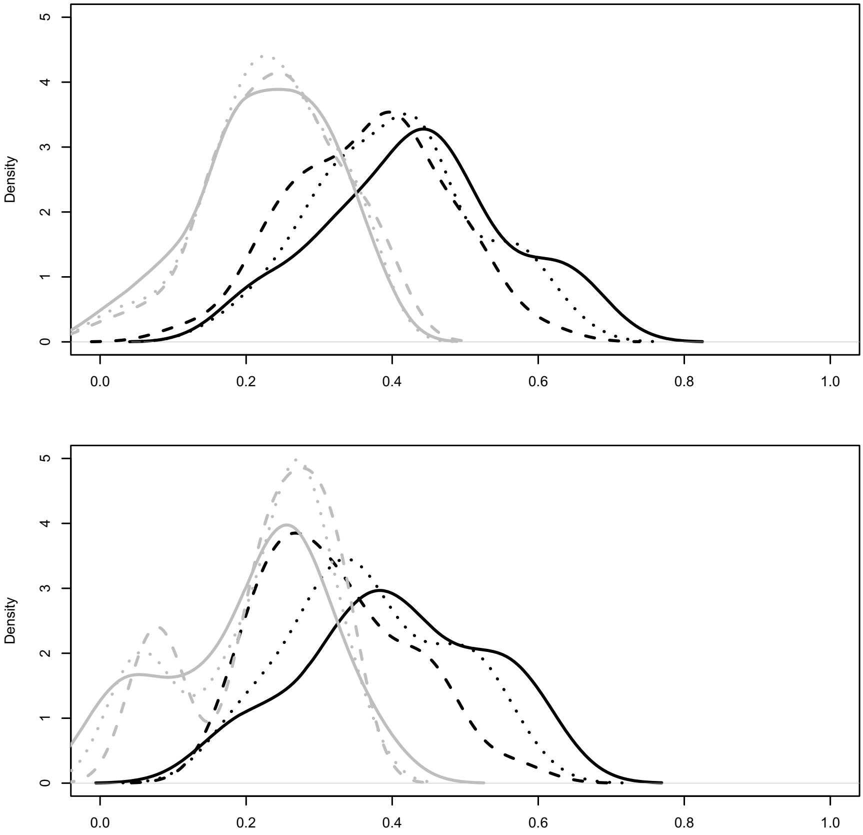
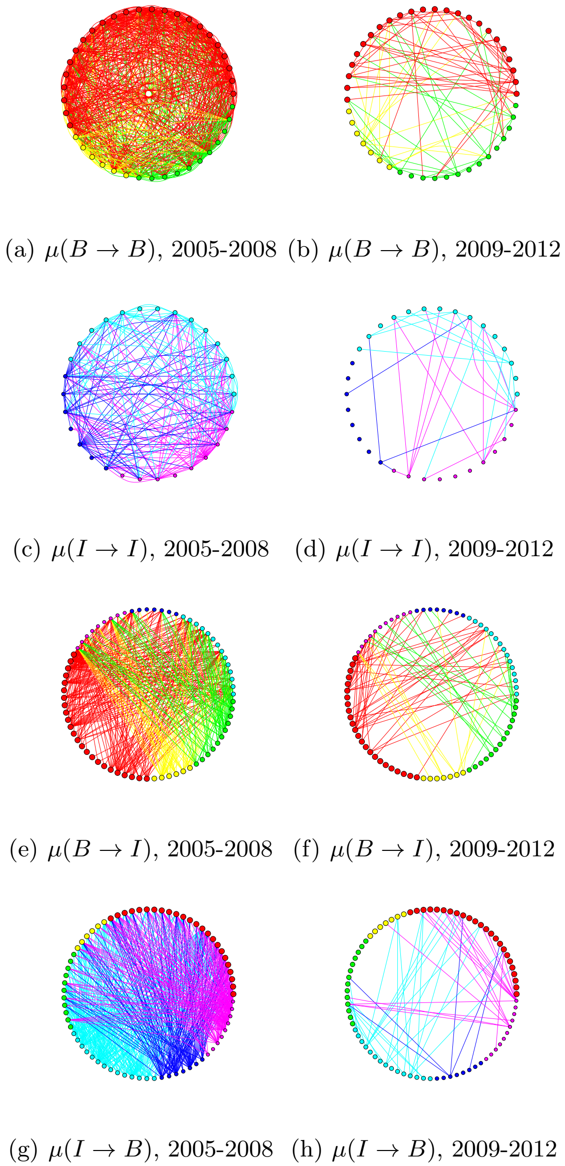


Fig. 1. The density curve of the contagion measure from a single financial institution, namely bank (*upper chart*) and insurance company (*lower chart*), to different sets: the bank sector (*solid*), the insurance sector (*dashed*), and both sectors (*dotted*). Black and gray colors refer to the before and after the crisis period, respectively.



(a) $\mu(B \to B)$, 2005-2008 (b) $\mu(B \to B)$, 2009-2012

(c) $\mu(I \to I)$, 2005-2008 (d) $\mu(I \to I)$, 2009-2012

(e) $\mu(B \to I)$, 2005-2008 (f) $\mu(B \to I)$, 2009-2012

(g) $\mu(I \to B)$, 2005-2008 (h) $\mu(I \to B)$, 2009-2012

Fig. 2. Extreme (> 0.5) asymmetric contagion measures within the sectors: Panels a and b refer to the bank sector B, while c and d the insurance sector I. Extreme (> 0.5) asymmetric contagion measures between the sectors: Panels e and f show the effects from bank to insurance sector, while g and f from insurance to bank sector. First column charts concern 2005-2008, while second column charts 2009-2012. The red, yellow, and green vertices stand for the large-, medium-, and small-sized banks, respectively; the cyan, blue, and purple vertices stand for the large-, medium-, and small-sized insurance companies, respectively.

4 Conclusions

We proposed the spatial contagion measure due to [5] for analyzing the connectedness from a single institution to one sector, and within and between the bank and insurance sectors, before and after the 2008 subprime crisis.

We found the contagion measures from single insurance companies to bank sector seem to be larger (or at least equal) than those from banks to the insurance sector. Moreover, while large-sized banks are more connected and affected by each other, insurance companies are more affected by medium- and small-sized companies. Finally, after the crisis, the contagion risk between financial institutions seems to reduce.

References

1. Adrian, T., Brunnermeier, M.: Covar. Staff Reports 348, Federal Reserve Bank of New York (2008), `http://ideas.repec.org/p/fip/fednsr/348.html`
2. Bernard, C., Brechmann, E., Czado, C.: Statistical assessments of systemic risk measures. In: Fouque, J.P., Langsam, J. (eds.) Handbook on Systemic Risk, pp. 165–179. Cambridge University Press, Cambridge (2013)
3. Billio, M., Getmansky, M., Lo, A., Pelizzon, L.: Econometric measures of connectedness and systemic risk in the finance and insurance sectors. J. Fin. Econ. 104(3), 535–559 (2012)
4. Durante, F., Foscolo, E.: An analysis of the dependence among financial markets by spatial contagion. Int. J. Intell. Syst. 28(4), 319–331 (2013)
5. Durante, F., Foscolo, E., Jaworski, P., Wang, H.: A spatial contagion measure for financial time series. Expert Syst. Appl. 41(8), 4023–4034 (2014)
6. Durante, F., Foscolo, E., Sabo, M.: A spatial contagion test for financial markets. In: Kruse, R., Berthold, M., Moewes, C., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Synergies of Soft Computing and Statistics. AISC, vol. 190, pp. 313–320. Springer, Heidelberg (2013)
7. Durante, F., Jaworski, P.: Spatial contagion between financial markets: a copula-based approach. Appl. Stoch. Models Bus. Ind. 26(5), 551–564 (2010)
8. Jaworski, P., Pitera, M.: On spatial contagion and multivariate GARCH models. Appl. Stoch. Models Bus. Ind. (2013) (in press) doi:10.1002/asmb.1977
9. Kolb, R. (ed.): Financial contagion: the viral threat to the wealth of nations. Wiley, Hoboken (2011)

Fuzzy Double Clustering: A Robust Proposal

Maria Brigida Ferraro and Maurizio Vichi

Department of Statistical Sciences, Sapienza University of Rome,
P.le A. Moro 5 - 00185 Rome, Italy
{mariabrigida.ferraro,maurizio.vichi}@uniroma1.it

Abstract. In this paper a robust fuzzy methodology for simultaneously clustering objects and variables is proposed. Starting from Double k-Means, different fuzzy generalizations for categorical multivariate data have been proposed in literature which are not appropriate for heterogeneous two-mode datasets, especially if outliers occur. In practice, in these cases, the existing fuzzy procedures do not recognize them. In order to overcome that inconvenience and to take into account a certain amount of outlying observations a new fuzzy approach with noise clusters for the objects and variables is introduced and discussed.

1 Introduction

Two-mode clustering consists in simultaneously clustering modes (e.g., objects, variables) of an observed two-mode data matrix. This idea arises to face with situations in which objects are homogeneous only within subsets of variables, while variables may be strongly associated only on subsets of objects (see Fig. 1). There are many practical applications presenting the above situations. For example, in DNA microarrays analysis groups of genes are generally co-regulated within subsets of samples and groups of samples share a common gene expression pattern only for some subsets of genes. In market basket analysis customers have similar preference patterns only on subsets of products and, vice-versa, classes of products are more frequently consumed and preferred by subgroups of customers. Other applications include biology, psychology, sociology and so on. By using a standard one-mode cluster analysis, the clusters of the objects are identified without considering that clusters of variables are present in the data. This problem can be overcome by simultaneously clustering the two modes. In this way all the information contained in heterogeneous datasets is completely taken into consideration. In case of heterogeneity outliers are likely to occur. These can be indicative of measurement errors or they are generated by a different data generation processes. An outlier can be an intermediate value between two clusters or can be far from all the remaining data. In two-mode dataset outliers can be due to different objects and variables generation mechanisms. In this work we propose a robust fuzzy two-mode clustering procedure in order to take into account different kinds of outliers. Starting from Double k-Means [9], we briefly recall some fuzzy extensions to categorical multivariate datasets, and we introduce a Fuzzy Double k-means with noise clusters. This procedure results to be robust to outliers.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_27

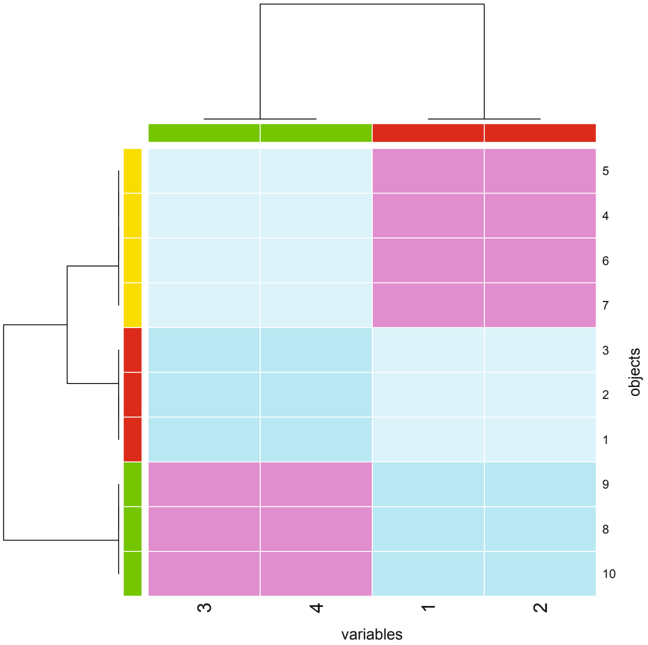


Fig. 1. Example of a 10 × 4 data matrix characterized by 3 × 2 homogeneous blocks

1.1 Notation

For the convenience of the reader the terminology used in this paper is listed here:

- n and p are the number of objects and of variables to be classified, respectively;
- $I = \{o_1, \cdots, o_i, \cdots, o_n\}$ is the set of n objects to be classified;
- $V = \{v_1, \cdots, v_j, \cdots, v_p\}$ is the set of p variables to be classified;
- $P = \{P_1, \cdots, P_g, \cdots, P_k\}$ is the partition of I into k classes, where P_g is the g-th class of P;
- $Q = \{Q_1, \cdots, Q_h, \cdots, Q_c\}$ is the partition of V into c classes, where Q_h is the h-th class of Q;
- $\mathbf{U} = [u_{ig}]$ is the $n \times k$ membership function matrix, assuming values in $[0, 1]$, specifying for each object o_i its membership to class P_g. Matrix $\mathbf{U}$, in this case, identifies a fuzzy classification of objects. When values of $\mathbf{U}$ are 1 or 0, i.e., $u_{ig} = 1$, object o_i belongs to P_g, while when $u_{ig} = 0$, object o_i does not belong to P_g. In this last case matrix $\mathbf{U}$ is binary and it identifies a hard classification of objects;
- $\mathbf{V} = [v_{jh}]$ is the $p \times c$ membership function matrix, assuming values in $[0, 1]$, specifying for each variable v_j its membership to class Q_h. Matrix $\mathbf{V}$, in this case, identifies a fuzzy classification of variables. In the hard case the values of $\mathbf{V}$ are 1 or 0;

- $\mathbf{Y} = [y_{gh}]$ is the $k \times c$ centroid matrix specifying the centroid of variable v_j in the class Q_h.

2 Double *k*-Means

Two-mode heterogeneous data are obtained by different generation mechanisms. They are characterized by different two-mode blocks that correspond to sub-matrices. Each sub-matrix contains objects homogeneous only on a subsets of variables and variables associated only on subsets of objects. For example, in Fig. 1 is reported a 10 × 4 matrix containing 3 × 2 blocks.

The standard or hard k-means [4] produces the partition of n objects into k clusters such that the sum of the within sum of squares of each cluster is minimized. A generalization of this approach in the two-mode case, the double k-means model, has been introduced in [9]. It is formally specified as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{Y}\mathbf{V}' + \mathbf{E}, \tag{1}$$

where matrix E is the error component matrix. The first term in (1) represents the information contained in the matrix $\mathbf{X}$ explained by the simultaneous classification of objects and variables. By using different constraints on the elements of the membership matrices $\mathbf{U}$ and $\mathbf{V}$, different classification structures can be defined.

In the standard or hard case the optimization problem is

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V},\mathbf{Y}} J_{DkM} = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{g=1}^{k} \sum_{h=1}^{c} \left(x_{ij} - y_{gh}\right)^2 u_{ig} v_{jh} \\ &\text{s.t.} \quad u_{ig}, v_{jh} \in \{0,1\}, \sum_{g=1}^{k} u_{ig} = 1, \sum_{h=1}^{c} v_{jh} = 1 \end{aligned} \tag{2}$$

Since the problem in (2) involves only binary variables u_{ig} and v_{jh} and a hard partition of objects and variables is required, double k-means can be solved using an alternating least squares (ALS) algorithm.

3 Fuzzy Clustering for Categorical Multivariate Data

In literature there are different proposals of fuzzy two-mode clustering for the specific case of categorical multivariate data. In a categorical multivariate dataset n individuals are described by a set of qualitative variables with p categories. These data are contained in tables, whose rows are the individuals and the columns are the categories. These are called cross-classification tables, contingency tables or in general co-occurrence matrices.

Since standard or fuzzy k-means type clustering algorithms (see, for example, [1]) are based on the distances from cluster centers (prototypes) to data points, it is not possible to consider them in this context. It is inappropriate to calculate those distances with respect to categorical data ([6]). Oh *et al.* [6] propose a generalization of fuzzy k-means [1] for categorical multivariate data. In details, the fuzziness is represented by an entropy regularization as in [5]. The optimization problem is defined as

$$\begin{aligned}
&\min_{\mathbf{U},\mathbf{V}} J_{FCCM} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{k} d_{ij}u_{ig}v_{jg} \\
&-T_u \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}\ln u_{ig} - T_v \sum_{j=1}^{p}\sum_{g=1}^{k} v_{jg}\ln v_{jg} \\
&\text{s.t.}\quad u_{ig}, v_{jg} \in [0,1], \sum_{g=1}^{k} u_{ig} = 1, \sum_{j=1}^{p} v_{jg} = 1,
\end{aligned} \tag{3}$$

where d_{ij} represents the co-occurrence of object i and category j. T_u and T_v are the degrees of fuzziness of the objects partition and of the variables partition, respectively. It is important to note that, for each i-th individual, the sum of the membership degrees of this individual to all the clusters and, for each g-th cluster, the sum of the membership degrees of all the categories to this cluster have to be equal to 1. The last constraint is different from that in (2). The membership degrees u_{ig} and v_{jg} have different constraints. In this case the optimization problem is minimized when only one variable in each cluster is completely relevant and the remaining ones are irrelevant. Hence, this turns out to be a "variable selection" procedure. The problem is solved by means of an iterative algorithm. Unfortunately, in presence of large numbers of individuals and categories, FCCM can imply numerical instabilities. In order to overcome that drawback, Fuzzy Codok was proposed by Kummamuru *et al.* [3]. It consists in considering as fuzzifier the Gini index rather than entropy in the objective function. As for fuzzy entropy, the Gini index is maximized when all u_{ig} and v_{jg} are equally distributed. The optimization problem is:

$$\begin{aligned}
&\min_{\mathbf{U},\mathbf{V}} J_{FCODOK} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{k} d_{ij}u_{ig}v_{jg} \\
&-T_u \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^2 - T_v \sum_{j=1}^{p}\sum_{g=1}^{k} v_{jg}^2 \\
&\text{s.t.}\quad u_{ig}, v_{jg} \in [0,1], \sum_{g=1}^{k} u_{ig} = 1, \sum_{j=1}^{p} v_{jg} = 1,
\end{aligned} \tag{4}$$

Since Fuzzy Codok allows the membership to take negative values, an additional step to perform clipping and renormalization is required in the optimization. Tjhi and Chen [7] propose to overcome that drawback by introducing a single term fuzzifier in the optimization problem. In details, the problem is formalized in the following way

$$\begin{aligned}
&\min_{\mathbf{U},\mathbf{V}} J_{FCC-STF} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{k} d_{ij}u_{ig}v_{jg} \\
&-T\sum_{i=1}^{n}\sum_{g=1}^{k}\sum_{j=1}^{p}\left((u_{ig}+v_{jg})-u_{ig}v_{jg}\right)^2 \\
&\text{s.t.}\quad u_{ig}, v_{jg} \in [0,1], \sum_{g=1}^{k} u_{ig}=1, \sum_{j=1}^{p} v_{jg}=1,
\end{aligned} \tag{5}$$

where T is the parameter of fuzziness. It has been proved that in this way the algorithm converges faster.

Unfortunately, the above methods are not appropriate for heterogeneous two-mode datasets.

4 Fuzzy Double *k*-Means

In this section a fuzzy approach for clustering heterogeneous two-mode datasets is proposed. The optimization problem J_{DkM} can be defined also for the fuzzy case when elements u_{ig} and v_{jh} assume values in $[0,1]$. In that case the double k-means can be solved using a sequential quadratic programming algorithm (see, for more details, [9]). Following the idea proposed by Bezdek [1], by introducing two parameters of fuzziness, m and l, the optimization problem can be written as

$$\begin{aligned}
&\min_{\mathbf{U},\mathbf{V},\mathbf{Y}} J_{FDkM} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{k}\sum_{h=1}^{c} (x_{ij}-y_{gh})^2 u_{ig}^m v_{jh}^l \\
&\text{s.t.}\quad u_{ig}, v_{jh} \in [0,1], \sum_{g=1}^{k} u_{ig}=1, \sum_{h=1}^{c} v_{jh}=1.
\end{aligned} \tag{6}$$

The parameters m and l represent the degrees of fuzziness of the objects and variables partitions, respectively.

By using Lagrangian multipliers the optimization problem is solved by means of derivatives with respect to the parameters and by means of an iterative algorithm. The updates of the membership degrees and the centroids are given by

$$u_{ig} = \frac{1}{\sum\limits_{g'=1}^{k}\left(\frac{\sum\limits_{j=1}^{p}\sum\limits_{h=1}^{c}(x_{ij}-y_{gh})^2 v_{jh}^l}{\sum\limits_{j=1}^{p}\sum\limits_{h=1}^{c}\left(x_{ij}-h_{g'h}\right)^2 v_{jh}^l}\right)^{\frac{1}{m-1}}}, \quad i=1,\cdots,n, \quad g=1,\cdots,k, \tag{7}$$

$$v_{jh} = \frac{1}{\sum\limits_{h'=1}^{c}\left(\frac{\sum\limits_{i=1}^{n}\sum\limits_{g=1}^{k}(x_{ij}-y_{gh})^2 u_{ig}^m}{\sum\limits_{i=1}^{n}\sum\limits_{g=1}^{k}\left(x_{ij}-h_{gh'}\right)^2 u_{ig}^m}\right)^{\frac{1}{l-1}}}, \quad j=1,\cdots,p, \quad h=1,\cdots,c, \tag{8}$$

$$y_{gh} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} u_{ig}^{m} v_{jh}^{l}}{\sum_{i=1}^{n} \sum_{j=1}^{p} u_{ig}^{m} v_{jh}^{l}}, \quad g = 1, \cdots, k, \quad h = 1, \cdots, c. \tag{9}$$

It is simple to prove that, when the number of variables clusters c is exactly equal to the number of variables p, we obtain $v_{jh} = 1$, $\forall j = 1, \cdots, p$, and the fuzzy double k-means corresponds to the fuzzy k-means.

5 Fuzzy Double *k*-Means with Noise

In many practical applications we have to take into account the presence of outliers. They are generated by a different mechanism with respect to the rest of the data and are not expected to belong to any two-mode block. In general outliers are significantly far from the closest block (see, for example, Fig. 2).

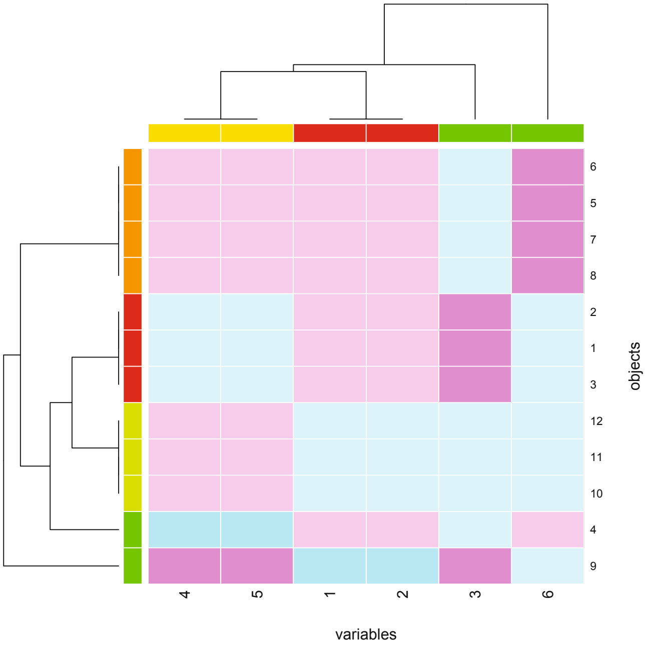


Fig. 2. Example of a 12 × 6 data matrix characterized by 2 Objects Outliers (4 and 9) and 2 Variables Outliers (3 and 6)

The performance of k-means or fuzzy k-means type algorithms is affected by outliers or noisy data. That problem is due to the constraints of the membership degrees. Each point is required to be assigned to one of the clusters in the standard case and in the fuzzy case the sum of the membership degrees is equal to 1. In this way also the outliers have to be assigned to the clusters and the results are strongly affected by these noisy points. In order to overcome that drawback Davé [2] introduced the concept of "Noise Cluster". The idea is to obtain an additional cluster containing all the outliers. It is assumed that the noise prototype has the same distance from all the points of the dataset. That distance is fixed a priori. We consider k good clusters for the objects and c for the variables. The $(k+1)$-th objects cluster and the $(c+1)$-th variables cluster are the noise ones. We fix a distance $\delta = (x_{ij} - y_{(k+1)(c+1)})^2$, for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Splitting the objective function (6) in two parts, one related to the good clusters and the other related to the noise clusters, we obtain

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\mathbf{Y}} J_{FDkMN} &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{k}\sum_{h=1}^{c} \left(x_{ij} - y_{gh}\right)^2 u_{ig}^m v_{jh}^l \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{p} \delta^2 \left(u_{i(k+1)}\right)^m \left(v_{j(c+1)}\right)^l \\ \text{s.t.} \quad & u_{ig} v_{jh} \in [0,1], \sum_{g=1}^{k+1} u_{ig} = 1, \sum_{h=1}^{c+1} v_{jh} = 1. \end{aligned} \tag{10}$$

Taking into account the above constraints, the optimization function becomes

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\mathbf{Y}} J_{FDkMN} &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{g=1}^{k}\sum_{h=1}^{c} \left(x_{ij} - y_{gh}\right)^2 u_{ig}^m v_{jh}^l \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{p} \delta^2 \left(1 - \sum_{g=1}^{k} u_{ig}\right)^m \left(1 - \sum_{h=1}^{c} v_{jh}\right)^l \end{aligned} \tag{11}$$

The performance of this approach has been investigated by means of simulation studies.

This approach can be used for different fuzzy double clustering algorithms, also those defined for categorical multivariate datasets.

6 Concluding Remarks

In this work we propose a robust fuzzy double k-means algorithm which includes as special case the standard fuzzy k-means. By introducing the concept of noise clusters in two-mode clustering, the results are not affected by objects and variables outliers.

References

1. Bezdek, J.C.: Cluster validity with fuzzy sets. Journal of Cybernetics 3, 58–73 (1974)
2. Davé, R.: Characterization and detection of noise in clustering. Pattern Recognition Letters 12, 657–664 (1991)
3. Kummamuru, K., Dhawale, A., Krishnapuram, R.: Fuzzy Co-clustering of Documents and Keywords. IEEE International Conf. on Fuzzy Systems 2, 772–777 (2003)

4. Mac Queen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, pp. 281–297 (1967)
5. Miyamoto, S., Mukaidono, M.: Fuzzy c-means as a regularization and maximum entropy approach. In: Proceedings of the Seventh International Fuzzy Systems Association World Congress (IFSA 1997), vol. II, pp. 86–92 (1997)
6. Oh, C.H., Honda, K., Ichihashi, H.: Fuzzy Clustering for Categorical Multivariate Data. In: Proc. of Joint Ninth IFSA World Congress and Twentieth NAFIPS International Conf., pp. 2154–2159 (2001)
7. Tjhi, W.-C., Chen, L.: Fuzzy Co-clustering of Web Documents. In: Proceedings of the 2005 International Conference on Cyberworlds, CW 2005 (2005)
8. Tjhi, W.-C., Chen, L.: Dual Fuzzy-Possibilistic Coclustering for Categorization of Documents. IEEE Transactions on Fuzzy Systems 17 (2009)
9. Vichi, M.: Double k-means clustering for simultaneous classification of objects and variables. In: Advances in Classification and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 43–52 (2001)

Sugeno Integral-Based Confidence Intervals for the Theoretical *h*-Index

Marek Gagolewski[1,2]

[1] Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland
gagolews@ibspan.waw.pl

[2] Faculty of Mathematics and Information Science, Warsaw University of Technology,
ul. Koszykowa 75, 00-662 Warsaw, Poland

Abstract. Sugeno integral-based confidence intervals for the theoretical h-index of a fixed-length sequence of i.i.d. random variables are derived. They are compared with other estimators of such a distribution characteristic in a Pareto i.i.d. model. It turns out that in the first case we obtain much wider intervals. It seems to be due to the fact that a Sugeno integral, which may be applied on any ordinal scale, is known to ignore too much information from cardinal-scale data being aggregated.

1 Introduction

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sequence of i.i.d. random variables with a common monotone strictly increasing c.d.f. F with support $\mathbb{I} = [0, \infty)$. The theoretical h-index, cf. [11], $\mathfrak{H}_n = \mathfrak{H}_n(X) \in (0, n)$ is a solution to:

$$1 - F(\mathfrak{H}_n) = \mathfrak{H}_n / n.$$

The theoretical h-index is a sample-size dependent location characteristic of a probability distribution. For example, if X follows a Pareto/Lomax distribution with $F(x) = 1 - 1/(1 + x)$, then $\mathfrak{H}_n = (\sqrt{4n+1} - 1)/2$.

Among estimators of $\mathfrak{H}_n$ we find the generalized Hirsch [12] index:

$$\widehat{h}_n(\mathbf{X}) = \bigvee_{i=1}^{n} X_{(n-i+1)} \wedge i = \max\left\{\min\{X_{(n)}, 1\}, \ldots, \min\{X_{(1)}, n\}\right\},$$

where $X_{(i)}$ denotes the ith smallest value in $\mathbf{X}$. Statistic $\widehat{h}_n$ is an OWMax [3,4] (and thus an OM3 [7]) operator corresponding to the Sugeno [14] integral of $\mathbf{X}$ with respect to the counting measure, see also [10,15]. What is important, it has already been shown (see [9] for the proof) that $\widehat{h}_n(\mathbf{X})/n$ is an asymptotically unbiased estimator of $\mathfrak{H}_n/n$.

It is well-known that the h-index, originally defined for a sample with elements in $\mathbb{N}_0$, has many fruitful applications, for example in bibliometrics [6], quality engineering [5] and information sciences [13]. However, still little is known on

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_28

the stochastic properties of such a measure. In [9,11] the properties of $\widehat{h}_n$ and other Sugeno integrals in an i.i.d. setting are considered, while in e.g. [1] its behavior in a more complex model is investigated. Moreover, in [8] a statistical test for the difference of h-indices in two Pareto-distributed random samples of equal lengths is derived and it turns out that such a tool has a very weak discriminatory power.

In this contribution we are interested in constructing Sugeno integral-based confidence intervals for the theoretical h-index, which is done in the section to follow. In Sec. 3 we provide some numeric examples for the Pareto distribution family. The obtained estimates are compared with different ones. It turns out that the $\widehat{h}_n$-based intervals are very wide, which is probably due to the fact that a Sugeno integral is known to ignore too much information from data. Finally, Sec. 4 concludes the paper.

2 Derivation of Sugeno Integral-Based Confidence Intervals

Fix n. Let $\Theta = (0, n)$ be a parameter space that induces an identifiable statistical model $(\mathbb{I}, \{\mathrm{Pr}_\theta : \theta \in \Theta\})^n$ in which for $X \sim \mathrm{Pr}_\theta$ we have $\theta = \mathfrak{H}_n(X)$ for all $\theta \in \Theta$, i.e. such that the theoretical h-index of X is equal to the value of parameter θ.

Definition 1. *Let $\alpha \in [0, 1]$. A random interval $\left(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X})\right)$ is called an $(1 - \alpha)$-confidence interval for parameter θ if:*

$$(\forall \theta \in \Theta) \quad \mathrm{Pr}_\theta \left(\underline{\theta}(\mathbf{X}) \le \theta \le \overline{\theta}(\mathbf{X})\right) \ge 1 - \alpha.$$

Of course, here we are interested in constructing the smallest confidence intervals which bounds are determined solely by the observed value of $\widehat{h}_n$. Additionally, we will assume a kind of symmetry of the intervals. The lower bound, $\underline{\theta}(\mathbf{X})$, will be defined via the smallest function $d_\alpha : (0, n) \to (0, n)$ such that for all $\theta \in (0, n)$ it holds

$$\mathrm{Pr}_\theta \left(\widehat{h}_n(\mathbf{X}) \le d_\alpha(\theta)\right) \ge 1 - \alpha/2.$$

Given the observed random sample realization $\mathbf{x}$ and $h = \widehat{h}_n(\mathbf{x})$, the lower bound will be determined by calculating $d_\alpha^{-1}(h) = \sup\{\theta : d_\alpha(\theta) \le h\}$. Thanks to such a setting we will have $\mathrm{Pr}_\theta(d_\alpha^{-1}(\widehat{h}_n(\mathbf{X})) \le \theta) \ge 1 - \alpha/2$.

On the other hand, the upper bound shall be given by the greatest function g_α such that

$$\mathrm{Pr}_\theta \left(\widehat{h}_n(\mathbf{X}) \ge g_\alpha(\theta)\right) \ge 1 - \alpha/2,$$

which is equivalent to $\mathrm{Pr}_\theta \left(\widehat{h}_n(\mathbf{X}) < g_\alpha(\theta)\right) \le \alpha/2$. This will provide us with $\mathrm{Pr}_\theta(\theta \le g_\alpha^{-1}(\widehat{h}_n(\mathbf{X}))) \ge 1 - \alpha/2$.

By [9, Lemma 2] we have:

$$\mathrm{Pr}_\theta(\widehat{h}_n(\mathbf{X}) \le h) = \mathcal{I}(\mathrm{Pr}_\theta(X \le h); n - \lfloor h \rfloor, \lfloor h \rfloor + 1),$$

where $\mathcal{I}(p; a, b)$ denotes the incomplete beta function of p with parameters a, b. We see that the c.d.f. of $\widehat{h}_n$ can be discontinuous even for continuous c.d.f. of X. Therefore,

$$\underline{\theta}(\mathbf{x}) = d_\alpha^{-1}(h) = \sup\left\{\theta : \mathcal{I}\left(\Pr{}_\theta(X < h); n - \lfloor h \rfloor, \lfloor h \rfloor + 1\right) \geq 1 - \alpha/2\right\},$$

and

$$\overline{\theta}(\mathbf{x}) = g_\alpha^{-1}(h) = \inf\left\{\theta : \mathcal{I}\left(\Pr{}_\theta(X \leq h); n - \lfloor h \rfloor, \lfloor h \rfloor + 1\right) \leq \alpha/2\right\}.$$

Unfortunately, in most cases the confidence interval bounds can only be calculated numerically.

3 Numerical Examples

For the sake of illustration let us consider the Pareto distribution family, $\mathcal{P}(k)$, with scale parameter $k > 0$. Such a distribution is sometimes used, cf. [11], in modeling empirical phenomena in the application scope of the h-index.

The cumulative distribution function of $X \sim \mathcal{P}(k)$ is defined by:

$$F(x) = 1 - \frac{1}{(x+1)^k} \quad (x \geq 0).$$

We have $\mathbb{E}\, X = 1/(k-1)$ for $k > 1$ and $\operatorname{supp} X = [0, \infty)$.

In order to guarantee that this family of distributions fits our statistical model's assumptions, we should introduce the following reparametrization. Let $\vartheta_n(k) = \mathfrak{H}_n(X)$ for $X \sim \mathcal{P}(k)$. Such a function may easily be calculated numerically with very good accuracy using some nonlinear root finding algorithm. Thus, we may consider $\mathcal{P}'(\theta) \equiv \mathcal{P}(\vartheta^{-1}(\theta))$, $\theta \in (0, n)$.

Figures 1 and 2 depict the 95%-confidence intervals bounds for $n = 10$ and 25, respectively. Note that the bounds are not continuous functions of $\widehat{h}_n$: they have jumps in points from the set $\{1, \ldots, n-1\}$. For example, for $n = 10$ and observed value of $\widehat{h}_n = 5$, we obtain an interval $(3.341, 7.779)$. On the other hand, for $\widehat{h}_n = 5^-$ we get $(2.840, 7.021)$.

We should also keep in mind that even though the obtained intervals are the smallest possible (at a confidence level of 95%), in fact the true probability of covering a theoretical h-index may sometimes be greater that 95%. This phenomenon, depicted in Figures 3 and 4, is of course consistent with the provided definition of a confidence interval. A similar behavior is observed e.g. for the Neyman-Clopper-Pearson (beta distribution-based, see [2]) confidence intervals for the probability of success in a Bernoulli experiment, cf. [16].

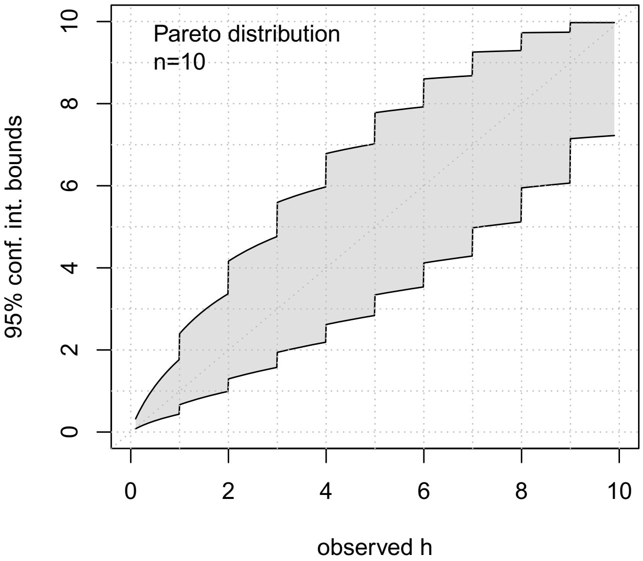


Fig. 1. Bounds for the Sugeno integral-based 95%-confidence intervals for the theoretical h-index; Pareto distribution family; $n = 10$

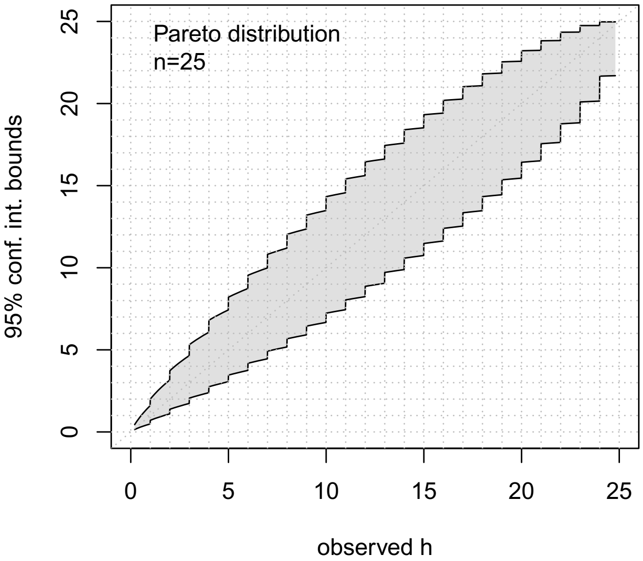


Fig. 2. Bounds for the Sugeno integral-based 95%-confidence intervals for the theoretical h-index; Pareto distribution family; $n = 25$

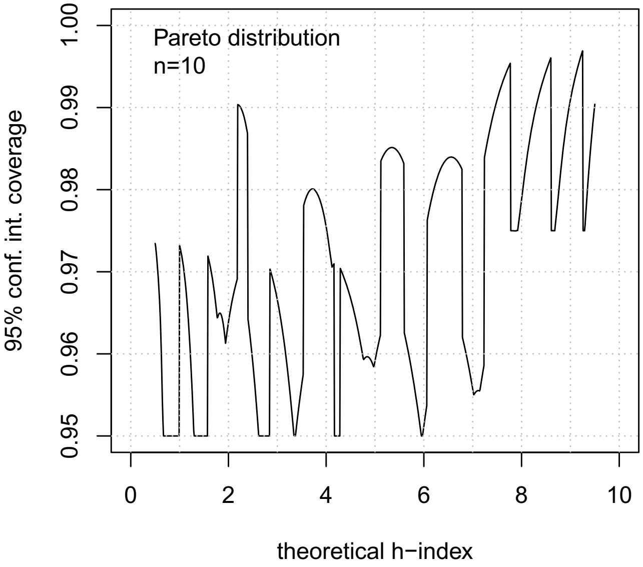


Fig. 3. Actual coverage of the true $\mathfrak{H}_n$ by Sugeno integral-based 95%-confidence intervals; Pareto distribution family; $n = 10$

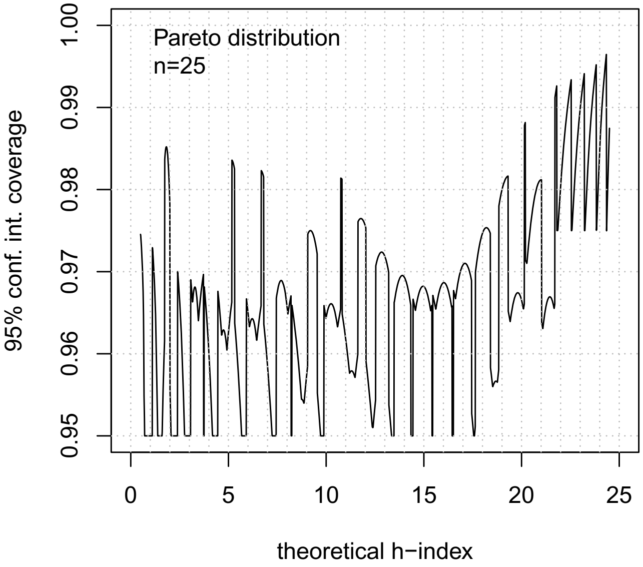


Fig. 4. Actual coverage of the true $\mathfrak{H}_n$ by Sugeno integral-based 95%-confidence intervals; Pareto distribution family; $n = 25$

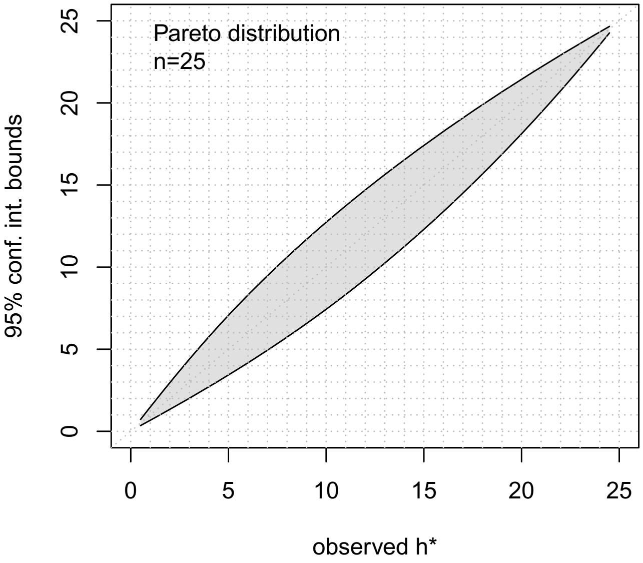


Fig. 5. Bounds for the $\widehat{h}_n^*$-based 95%-confidence intervals for the theoretical h-index; Pareto distribution family; $n = 25$

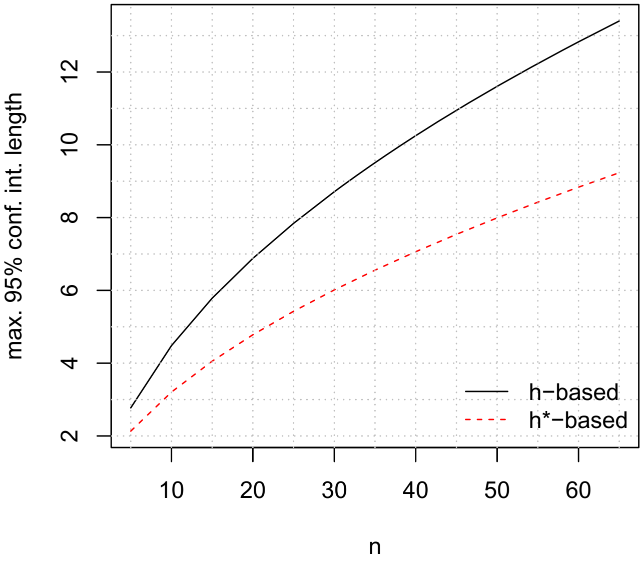


Fig. 6. Maximal widths of Sugeno integral- and $\widetilde{h}_n^*$-based 95-% confidence intervals for the theoretical h-index as a function of sample size n; Pareto distribution family

Comparison to other estimates. It might easily be shown that for $(X_1, \ldots, X_n)$ i.i.d. $\mathcal{P}(k)$ the statistic

$$\widehat{k}_n^*(\mathbf{X}) = (n-1) / \sum_{i=1}^{n} \log(1 + X_i)$$

is an unbiased and consistent estimator of k. What is more, $\sum_{i=1}^{n} \log(1 + X_i) \sim \Gamma(n, k)$.

We may thus try using $\widehat{h}_n^* = \vartheta_n(\widehat{k}_n^*)$ as an estimator of $\mathfrak{H}_n$. Numerical results indicate that $\widehat{h}_n^*/n$ may only be asymptotically unbiased estimator of $\mathfrak{H}_n/n$. By the above-mentioned fact, if $(X_1, \ldots, X_n)$ i.i.d. $\mathcal{P}'(\vartheta(k))$, then

$$\Pr{}_{\vartheta(k)}(\widehat{h}_n^*(\mathbf{X}) \le h) = 1 - G_{n,k}\left(\frac{n-1}{\vartheta^{-1}(h)}\right),$$

where $G_{n,k}$ is the c.d.f. of the gamma distribution $\Gamma(n, k)$. This time, such an estimator has a continuous distribution.

A $\widehat{h}_n^*$-based $(1-\alpha)$-confidence interval may be derived in a manner similar (but much simpler due to continuity) to the previously considered one. It is a random interval $(\underline{\theta}^*(\mathbf{X}), \overline{\theta}^*(\mathbf{X}))$ such that $\underline{\theta}^*(\mathbf{X}) = d_\alpha^{-1*}(h)$ and $\overline{\theta}^*(\mathbf{X}) = g_\alpha^{-1*}(h)$ for which it holds

$$\begin{aligned} \Pr{}_{d_\alpha^{-1*}(h)}(\widehat{h}_n^*(\mathbf{X}) \le h) &= \alpha/2, \\ \Pr{}_{g_\alpha^{-1*}(h)}(\widehat{h}_n^*(\mathbf{X}) \le h) &= 1 - \alpha/2. \end{aligned}$$

Again, these equations may be solved numerically with a nonlinear root finder. This time we obtain a confidence interval which is exactly at a confidence level of $1 - \alpha$ for each θ.

Figure 5 depicts $\widehat{h}_n^*$-based 95%-confidence interval bounds for $n = 25$. We observe that they are of smaller length than those presented in Figure 2. Moreover, interval lengths for different sample sizes are given in Figure 6. We note that $\widehat{h}_n^*$ are better quality estimates than the Sugeno integral-based ones.

4 Conclusions

In this paper we derived Sugeno integral-based confidence intervals for the theoretical h-index, which is a location-type characteristic of a probability distribution. Large widths of the Sugeno integral-based intervals for a sample from the Pareto distribution family may possibly be due to the fact that this aggregation method is known not to utilize “full information” in input data. For example, for $n = 6$, $\widehat{h}_n(\mathbf{x}) = 3$ is obtained for $\mathbf{x} = (3, 3, 3, 0, 0, 0)$ as well as for $\mathbf{x} = (\infty, \infty, \infty, 3, 3, 3)$.

Taking into account the close relationship between confidence intervals and statistical hypothesis tests, the presented results are consistent with conclusions of [8]: the nature of Sugeno integral allows its application on any ordinal scale, but the prize we are paying for its robustness is the lack of good performance for cardinal scales.

Acknowledgments. The author would like to thank the anonymous reviewers for comments and suggestions.

References

1. Burrell, Q.L.: Hirsch's h-index: A stochastic model. Journal of Informetrics 1, 16–25 (2007)
2. Clopper, C., Pearson, E.: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26, 404–413 (1934)
3. Dubois, D., Prade, H.: Semantics of quotient operators in fuzzy relational databases. Fuzzy Sets and Systems 78(1), 89–93 (1996)
4. Dubois, D., Prade, H., Testemale, C.: Weighted fuzzy pattern matching. Fuzzy Sets and Systems 28, 313–331 (1988)
5. Franceschini, F., Maisano, D.A.: The Hirsch index in manufacturing and quality engineering. Quality and Reliability Engineering International 25, 987–995 (2009)
6. Franceschini, F., Maisano, D.A.: Structured evaluation of the scientific output of academic research groups by recent h-based indicators. Journal of Informetrics 5, 64–74 (2011)
7. Gagolewski, M.: On the relationship between symmetric maxitive, minitive, and modular aggregation operators. Information Sciences 221, 170–180 (2013)
8. Gagolewski, M.: Statistical hypothesis test for the difference between hirsch indices of two pareto-distributed random samples. In: Kruse, R., Berthold, M., Moewes, C., Gil, M.A., Grzegorzewski, P., Hryniewicz, O., et al. (eds.) Synergies of Soft Computing and Statistics. AISC, vol. 190, pp. 359–367. Springer, Heidelberg (2013)
9. Gągolewski, M., Grzegorzewski, P.: S-statistics and their basic properties. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano, M.A., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. AISC, vol. 77, pp. 281–288. Springer, Heidelberg (2010)
10. Gagolewski, M., Mesiar, R.: Monotone measures and universal integrals in a uniform framework for the scientific impact assessment problem. Information Sciences 263, 166–174 (2014)
11. Glänzel, W.: On some new bibliometric applications of statistics related to the h-index. Scientometrics 77(1), 187–196 (2008)
12. Hirsch, J.E.: An index to quantify individual's scientific research output. Proceedings of the National Academy of Sciences 102(46), 16569–16572 (2005)
13. Hovden, R.: Bibliometrics for internet media: Applying the h-index to YouTube. Journal of the American Society for Information Science and Technology 64(11), 2326–2331 (2013)
14. Sugeno, M.: Theory of fuzzy integrals and its applications. Ph.D. thesis, Tokyo Institute of Technology (1974)
15. Torra, V., Narukawa, Y.: The h-index and the number of citations: Two fuzzy integrals. IEEE Transactions on Fuzzy Systems 16(3), 795–797 (2008)
16. Zieliński, R.: Confidence intervals for proportions (Przedziały ufności dla frakcji). Matematyka Stosowana 10, 51–68 (2009) (in Polish)

Using Changes in Distribution to Identify Synchronized Point Processes

Christian Braune, Stephan Besecke, and Rudolf Kruse

Otto-von-Guericke-University of Magdeburg
Universitätsplatz 2, D-39106 Magdeburg, Germany
{christian.braune,rudolf.kruse}@ovgu.de, stephan.besecke@st.ovgu.de

Abstract. In neurobiology the analysis of spike trains is of particular interest. Spike trains can be seen as point processes generated by neurons emitting signals to communicate with other neurons. According to Hebb's seminal work on neural encoding information is processed in the brain in ensembles of neurons that reveal themselves by synchronized behaviour. One of the many competing hypotheses to explain this synchrony is the spike-time-synchrony hypothesis. The relative timing of spikes emitted by different neurons should explain the processing of information. In this paper we present a novel method to decide for each single neuron whether it is part of (at least) one assembly by analyzing changes in the distribution of spiking patterns.

Keywords: point processes, distribution change, spike train analysis.

1 Introduction

Point processes occur in many situations such as the number of incoming phone calls in a call-center, sensor readings or as abstraction of biological processes such firing neurons. If a neuron is excited by collecting enough neuro-transmitters via its dendritical connections it releases an electrical discharge which travels along its axon and initiates the release of neuro-transmitters itself. This electrical discharge can be recorded by micro-electrode arrays. If the recorded electrical potential exceeds a certain threshhold, a spike can be detected and its point in time is recorded. The list of such points is called a spike train. Spike trains may be the result of *in-vivo* or *in-vitro* measurements of neuronal activity. Figure 1 shows two set of spike trains. While the right plot only shows random noise, the left one includes one assembly of 20 neurons that fire together more often (not always) than can be expected just by chance. Even the trained eye will only recognize noise in both plots. While the process of single neurons firing is fairly well understood and developments such as the Hudgkin-Hoxley model [8] allow to precisely model the electrical potential emitted by a single neuron, and the interactions between different regions of the brain as well, the interactions between larger groups of single neurons is still not understood. Many different theories have been developed so far that try to explain the cooperation of neurons, most

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_29

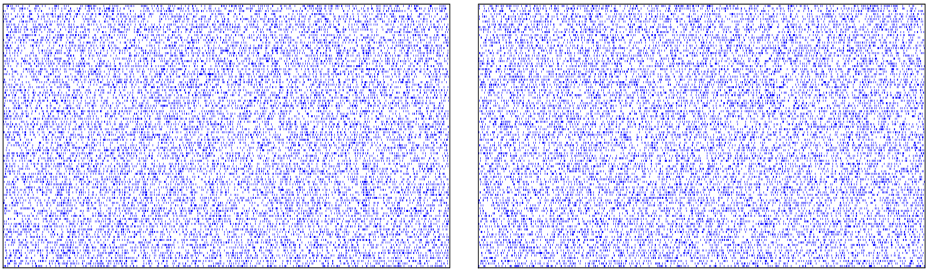


Fig. 1. Two sets of parallel spike trains, right one shows only random noise, left one has an ensemble of 20 neurons inserted by injecting coincident spike events (1ms precision) (see also: [6,5,1]).
x-axis: Time (10s), binned with 1ms precision, y-axis: spike train id

remarkably the neuron assembly theory introduced by Hebb [7]. Within this theory neurons are organized in so-called assemblies that exhibit themselves by increased synchronous behaviour. Such behaviour might be a simultaneous increase of their firing rate whenever a stimulus is presented or spikes that are emitted at (roughly) the same time. Since the recording of hundreds of spike trains in parallel is possible nowadays through the use of multielectrode arrays (MEAs), methods for the analysis of such relatively large data sets are needed.

In this work we will investigate further into how neurons that belong to an assembly can be efficiently distinguished from those that do not. This task can be seen as a binary classification problem where all data points are unlabeled at the beginning. The two classes we may assign to a point are *assembly* or *noise* neuron, indicating that the neuron either is part of an assembly or not. Common approaches try to focus on assigning the first label correctly and each noise neuron that is labeled as an assembly neuron can be seen as false positives. Thus, the common null hypothesis is: All neurons are independent of each other. For this work we will investigate how reversing this task influences the classification accuracy. Thus our null hypothesis is: All neurons are dependent of each other. This leads to our goal of identifying those neurons that are *not* part of any assembly.

In the following section we will review some methods that are currently used to identify neuronal assemblies and how they differ from our proposed method. In Section 3 we will present our approach and evaluate it in Section 4. The paper ends with a conclusion and an outlook into some future applications in Section 5.

2 Related Work

In [4] a first algorithm for detecting neuronal assemblies has been presented. It relies on the pairwise comparison of binned spike train data and a χ^2 test for independence for these. Two spike trains for which the null hypothesis of independence could be rejected are subsequently merged and the process is repeated until no further pair could be found. The order in which spike trains where merged determines a graph structure in which assemblies are revealed as cliques.

Lower-level algorithms only determine whether an assembly is present in the given data or not [9,11,12] or they test whether a single neuron belongs to such an assembly [1]. Higher-level algorithms are able to identify the assembly structure (such as in [10,2,3]) even under difficulties such as *selective participation* (i.e. not every neuron that belongs into an assembly takes part in a coincidence) or *temporal imprecision* (spikes are not perfectly aligned across different spike trains).

Similar to our approach are the surrogate-based algorithms (e.g. [1]) in which certain properties of a spike train are retained while others are purposely destroyed. Whenever a synchronous pattern (such as the number of coincidences with other spike trains) occurs in the independently generated surrogate it might be explained by pure randomness and the likelihood of a true synchronous behavior is reduced. By simply counting the number of times a synchronous pattern emerges or a more extreme value of any test statistic used is observed in the surrogate data and dividing by the number of surrogates generated, an empirical p-value can be gained.

3 Method

First we will start by formalizing how we understand a spike train to help analyse the spike trains' population's structure. Thus, let $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ be an ordered set with $p_0 \leq t_1 < \ldots < t_i < \ldots < t_n \leq p_1, t_i \in \mathbb{R}, \forall 1 < i < m$ and $m = \|\mathcal{T}\|$. Such a set we call spike train and each $t_i \in \mathcal{T}$ represents the time, when a spike was recorded as being emitted from a neuron within a recording of length $p_1 - p_0 = p$. Usually each t_i is either given in seconds or milliseconds.

A set of *parallel* spike trains is a set of spike trains $\mathcal{S} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$ if for every $\mathcal{T} \in \mathcal{S}$ p_0 and p_1 are identical, i.e. the recordings of the spike trains happened at the same time (e.g. by means of a MEA).

A *binned* spike train is a vector over $\{0,1\}^k$ whereas k is the number of non-intersecting, fixed-length windows that can be laid over the period p, i.e. $k = \lceil p/w \rceil$ for w being the window length. Usually windows are chosen to be $1ms$ long. The j-th compnent of a binned spike train $T \in \{0,1\}^k$ is equal to 1 if and only if there exists a spike in the frame represented by this component. I.e. $T[j] = 1 \leftrightarrow \exists t \in \mathcal{T} : j \cdot w \leq t < (j+1) \cdot w$, where $T[j]$ refers to the j-th component of the vector T. Such binning might lead to a loss of information since the real number of spikes might be higher than the number of components in the vector which are one (this is the case, if more than one spike lie in the same time bin and is referred to as *clipping*).

With this, every component of a binned spike train can be seen as a simple Bernoullli experiment. The probability for each spike train to yield a one for a component can be estimated as $\hat{p_T} = \frac{1}{k}\sum_{j=1}^{k} T\left[j\right]$ by the number of time bins which contain at least one spike over the total number of time bins. As such, we can count how many bins exist in which no, only one, two, etc. spike trains were active. If the spike trains were truly independent this distribution should follow a binomial distribution [1]. Assemblies, which activate synchronously, yield a higher amount of activations around their respective assembly size while reducing the number of activations for lower numbers. A simple χ^2 test could therefore reveal the presence of an assembly fairly easily [5]. What we are interested in is to answer the question whether each single spike belongs to an assembly or not.

With an estimate for the overall firing probability of a spike train across the whole spike train population

$$\hat{p} = \frac{1}{n \cdot k} \sum_{i=1}^{n} \sum_{j=1}^{k} T_i \left[j\right] \tag{1}$$

we can give the distribution of expected numbers of spike train activations per time bin (spike pattern) as the probability mass function of a binomial distribution with paramter $\hat{p}$:

$$f(k; n, \hat{p}) = \binom{n}{k} \hat{p}^k (1 - \hat{p})^{n-k}. \tag{2}$$

Figure 2 shows how the distribution of spike patterns changes if an assembly is present or not. Spike trains that are part of an assembly contribute more to the dent in the empirical distribution than truly independent neurons do. As such, replacing them by randomly generated neurons with similar characteristics (i.e. the firing probability or number of spikes stays the same) should change the distribution towards the binomial distribution estimate. On the other hand, replacing a neuron that is not part of an assembly should at most slightly change the distribution. Based upon this observation we can derive a test for classifying each single spike train by replacing it with a spike train that is generated independently from all other spike trains. By looking at the distribution change induced by replacing the spike train we can reason whether or not the neuron in question belongs into an assembly or is truly independent. Hence, a significantly lower χ^2 value would indicate that the original, replaced spike train contributed to the difference to a truly independent distribution. Thus, we calculate the ratio between the old χ^2 value and the new χ^2 value as $\rho_i = {\chi_i^2}/{\chi_{all}^2}$ (if the ith spike train was replaced) to make the assessment of the difference between assembly and non-assembly spike trains independent of the size of a possibly existing assembly. Values of ρ_i that are significantly smaller than one indicate a *more* independent distribution, thus hinting at a synchronized spike train having been replaced.

If – on the other hand – this ratio is not substantially smaller than one it indicates the original data and the surrogate are both truly independent. Since

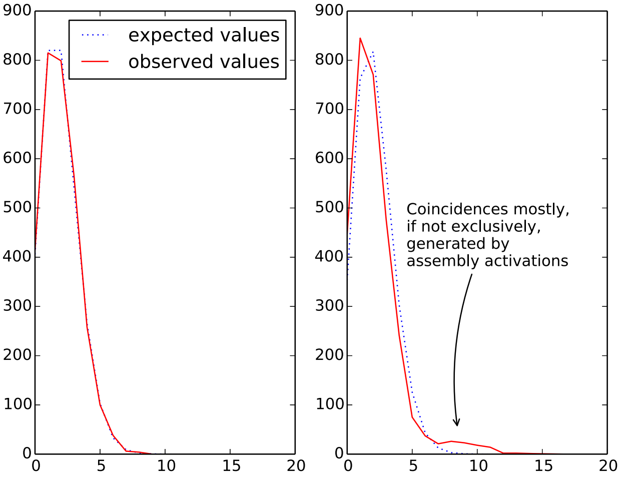


Fig. 2. Left: Theoretical distribution with estimated $\hat{p}$ (dotted line) and observed/empirical distribution for a set of independent processes.
Right: Theoretical distribution with estimated $\hat{p}$ (dotted line) and observed/empirical distribution for a set with an assembly.

we want to find out which spike trains are independent of the rest, we can exclude spike trains that led to a very small value of ρ_i. Values close to one indicate that no significant change happened and thus the replaced spike train might be as independent as the one it has been replaced with. The effects of either replacement can be seen in Figure 3.

A test solely based on such a test statistic that was used to identify the synchronized spike trains, would indeed yield a lot of false positive results. Let us consider two independent spike trains. Any number of coincidences between those two that is large enough will distort the ideal distribution and lead to high χ^2 values. Replacing those independent spike trains by other independent spike trains will therefor produce small values for their respective ratios ρ_i. Since we are not testing for the dependent trains but want to identify the independent ones, we can consider these cases as *false negatives*. Thus, we implemented a second identfication phase in which all processes that could not be marked as independent with absolute certainty form a new, smaller population and the test is repeated for these spike trains. Since usually a lot of the processes can already be marked in the first step, the proportion of independent spike trains is smaller and the identification becomes easier. All points that could already be marked as ones in the first step (by exceeding a empirically derived threshold) were excluded from the data set and only those points where the classification result remained unclear were tested again in a second tagging step.

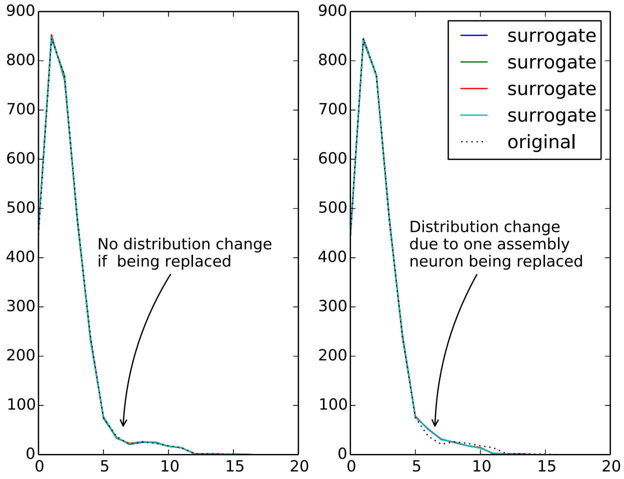


Fig. 3. Left: Original distribution with changes indicated by solid lines if an independent spike train is replaced. No distribution change is recognizable.
Right: Original distribution with changes indicated by solid lines if an assembly spike train is replaced. a distribution change is obvious.

4 Evaluation

To show the effectiveness of our method we generated several sets of artificial parallel spike trains. This has the advantage that we have a ground truth to validate our results with. Spike trains were modeled as Poisson processes of $10s$ length and with an average firing rate of $20Hz$. Assemblies of different sizes were injected and their spikes copied into the otherwise independently generated spike trains with probabilities of either 1.0, 0.8 or 0.6. Spike trains which could still not be labeled explicitly were given the label $'?'$.

Figures 4,5 and 6 show the results of 1000 tests performed each for a data set of 100 processes of which 10, 20 or 30 were synchronized. The number of false positive classifications naturally increases with decreasing copy probabilities, but always stays below 0.02% for results obtained from the second phase. This comes at the cost that several spike trains that were originally independent were tagged as synchronized since they fired more often than the average spike train together with other processes (eventually the assembly). Thus the number of false positives is relatively high at constantly around 6%. Also, the algorithm may abstain classifying a point at all, which occurs with the same likelihood. The results indicate that we can increase the certainty of a classification of a spike train as *independent* but only at the cost of having more points labeled as ?, thus decreasing type II errors.

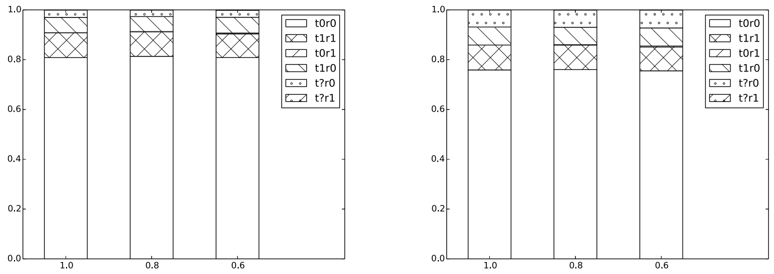


Fig. 4. Test results for first (left) and second tagging phase for 1000 samples of 100 spike trains each, with 10 synchronized processes injeected

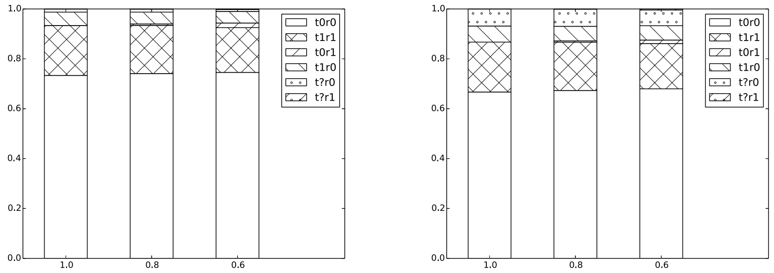


Fig. 5. Test results for first (left) and second tagging phase for 1000 samples of 100 spike trains each, with 20 synchronized processes injeected

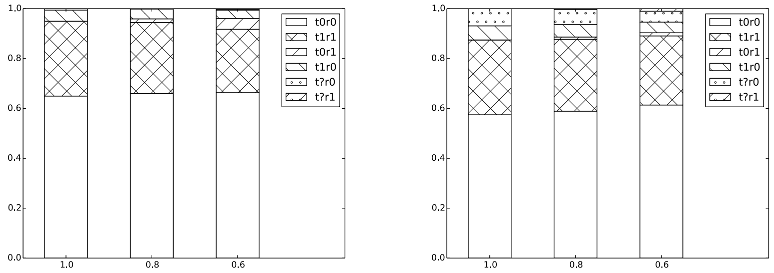


Fig. 6. Test results for first (left) and second tagging phase for 1000 samples of 100 spike trains each, with 30 synchronized processes injeected

5 Conclusion and Future Work

In this paper we showed that the identification of synchronized and independent point processes can be achieved by a method that uses changes in the empirical distribution of spike patterns as decision criterion. Classifications can be made for the whole population or for a single process only which enables us to investigate active learning scenarios for assembly detection.

Especially in conjunction with tagging algorithms that are testing for the dependent processes we can generate nearly certain labels for either class and only if the methods contradict each other we omit any label and can use this in semi-supervised learning.

Although there is spike train data available that has been obtained from real neurons, and our algorithm could analyze those data sets the validity of any result cannot be verified. Neural information encoding is still not well enough understood to actually verify the results we are having and we can merely present the tools needed to perform the analysis.

References

1. Berger, D., Borgelt, C., Louis, S., Morrison, A., Grün, S.: Efficient identification of assembly neurons within massively parallel spike trains. Computational Intelligence and Neuroscience 1 (2010)
2. Borgelt, C., Braune, C.: Prototype construction for clustering of point processes based on imprecise synchrony. In: 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2013). Atlantis Press (2013)
3. Braune, C., Borgelt, C., Kruse, R.: Behavioral clustering for point processes. In: Tucker, A., Höppner, F., Siebes, A., Swift, S. (eds.) IDA 2013. LNCS, vol. 8207, pp. 127–137. Springer, Heidelberg (2013)
4. Gerstein, G.L., Perkel, D.H., Subramanian, K.: Identification of functionally related neural assemblies. Brain Research 140(1), 43–62 (1978)
5. Grün, S., Abeles, M., Diesmann, M.: Impact of higher-order correlations on coincidence distributions of massively parallel data. In: Marinaro, M., Scarpetta, S., Yamaguchi, Y. (eds.) Dynamic Brain - from Neural Spikes to Behaviors. LNCS, vol. 5286, pp. 96–114. Springer, Heidelberg (2008)
6. Grün, S., Diesmann, M., Aertsen, A.: Unitary events in multiple single-neuron spiking activity: I. detection and significance. Neural Computation 14(1), 43–80 (2002)
7. Hebb, D.O.: The organization of behavior: a neuropsychological theory. Wiley (1949)
8. Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of Physiology 117(4), 500 (1952)
9. Louis, S., Borgelt, C., Gruen, S.: Complexity distribution as a measure for assembly size and temporal precision. Neural Networks 23(6), 705–712 (2010)
10. Picado-Muiño, D., Borgelt, C., Berger, D., Gerstein, G., Grün, S.: Finding neural assemblies with frequent item set mining. Frontiers in Neuroinformatics 7 (2013)
11. Staude, B., Gruen, S., Rotter, S.: Higher-order correlations in non-stationary parallel spike trains: Statistical modeling and inference. Frontiers in Computational Neuroscience 4(16) (2010)
12. Staude, B., Rotter, S., Gruen, S.: Cubic: Cumulant based inference of higher-order correlations in massively parallel spike trains. Journal of Computational Neuroscience 29(1-2), 327–350 (2010)

Lift Measure for Fuzzy Association Rules

Michal Burda

Institute for Research and Applications of Fuzzy Modeling,
Centre of Excellence IT4Innovations, University of Ostrava
30. dubna 22, 701 03 Ostrava, Czech Republic
michal.burda@osu.cz

Abstract. The aim of this paper is to provide a correct definition of lift measure for fuzzy association rules, to study some of it's interesting mathematical properties, and to provide an algorithm for fast computation of fuzzy lift during the process of fuzzy association rules mining.

Keywords: fuzzy association rules, lift, algorithm.

1 Introduction

Searching for association rules is a broadly discussed, developed and accepted data mining technique [9,2]. An association rule is an expression $X \rightarrow Y$, where antecedent X and consequent Y are conditions – the former usually in the form of elementary conjunction. Such rules are usually interpreted as implication statement "if X is satisfied, then Y is true very often, too". Two traditional measures of intensity of an association rule are often used, *support* and $confidence$. An objective is to find rules with *support* and $confidence$ above some user-defined thresholds.

Searching for association rules fits particularly well on binary or categorical data and many has been written on that topic [9,2,3,18]. For association analysis on numeric data, a prior discretization is proposed e.g. by Srikant et al. [17]. Another alternative is to take advantage of the fuzzy sets theory.

The use of fuzzy sets in connection with association rules has been motivated by many authors (see [11] for recent overview). Fuzzy association rules are appealing also because of the use of vague linguistic terms such as "small", "very big" etc. [7,14]. Fuzzy rule mining algorithms are usually based on well-known algorithms developed for binary data such as *Apriori* [2,16] or *FP-tree* [10].

In this paper, we focus on *lift*, a measure of intensity of a relationship among conditions of a rule. Lift was initially developed for non-fuzzy (i.e. "crisp") association rules and it is probably firstly described in [4] under its original name "interest". Lift has been well studied for association rules on binary data [4,13,8] Moreover, a nice overview of many other crisp rule measures provides [12].

As quite many was written about lift for crisp association rules, not so much has been done on lift for fuzzy rules. Some authors believe the generalization of lift for fuzzy rules is as trivial as substituting crisp terms with analogous fuzzy terminology inside of a crisp lift definition – see e.g. [5,15]. Unfortunately, as

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_30

discussed in this paper, such over-simplication may lead to erroneous outputs. In order to preserve some nice mathematical properties of lift, one must take care to define fuzzy lift appropriately.

In Section 2, a brief theoretical background for both binary and fuzzy association rules is provided. Section 3 gives a correct definition of fuzzy lift and discusses some of its interesting mathematical properties. The difference of lift w.r.t. underlying t-norm is also highlighted there. Section 4 introduces a fast algorithm for computation of fuzzy lift together with it's time and space complexity analysis. Finally, the last section summarizes the achieved goals and draws possible directions of future research.

2 Theoretical Background

2.1 Binary Association Rules

Let $\mathcal{O} := \{o_1, o_2, \ldots, o_n\}$, $n > 0$, be a finite set of objects and $\mathcal{A} := \{a_1, a_2, \ldots, a_m\}$, $m > 0$ be a finite set of attributes (features). Each attribute can be considered as a logical predicate: $a_i(o_j)$ is true (i.e. $a_i(o_j) = 1$), resp. false (i.e. $a_i(o_j) = 0$), if the i-th attribute applies, resp. does not apply, to object o_j. For a subset $X \subseteq \mathcal{A}$ of attributes, let us define a new predicate of a logical conjunction of the attributes contained in X:

$$X(o_j) \;:\equiv\; \forall a_i \in X : a_i(o_j). \tag{1}$$

An association rule is a formula $X \rightharpoonup Y$, where $X \subset \mathcal{A}$ is an *antecedent*, $Y \subset \mathcal{A}$ is a *consequent* and $X \cap Y = \emptyset$. Both X and Y are also called *itemsets.* Please consider the following rule as an example: $\{\text{tequila}, \text{salt}\} \rightarrow \{\text{lemon}\}$.

There are defined many quality measures for the association rules [12]. Among them, the most common are *support* and *confidence* [2,1]:

$$\text{supp}(X) := \frac{\big|\{o \in \mathcal{O} \mid X(o)\}\big|}{n}, \tag{2}$$

$$\text{supp}(X \rightharpoonup Y) := \text{supp}(X \cup Y), \tag{3}$$

$$\text{conf}(X \rightharpoonup Y) := \frac{\text{supp}(X \rightharpoonup Y)}{\text{supp}(X)}, \tag{4}$$

where $n = |\mathcal{O}|$.

If $\mathcal{O}$ is a random sample, then choosing o into $\mathcal{O}$ is a random event. Then a random event X (resp. Y) may be defined on the basis of the truth value of the predicate $X(o)$ (resp $Y(o)$). Then support $\text{supp}(X)$ (resp. $\text{supp}(Y)$) is an estimate of a probability $\mathrm{P}(\mathsf{X})$ (resp. $\mathrm{P}(\mathsf{Y})$). (Please note that $\text{supp}(X \cup Y)$ equals to the probability of events X and Y occuring together, i.e. $\mathrm{P}(\mathsf{X} \wedge \mathsf{Y})$.) Confidence $\text{conf}(X \rightharpoonup Y)$ is then an estimate of conditional probability $\mathrm{P}(\mathsf{Y}|\mathsf{X})$.

Lift (originally called *interest* in [4]) is defined as:

$$\text{lift}(X \rightharpoonup Y) := \frac{\text{conf}(X \rightharpoonup Y)}{\text{supp}(Y)}. \tag{5}$$

If X and Y are stochastically independent, $\mathrm{P}(\mathsf{Y}|\mathsf{X}) = \mathrm{P}(\mathsf{Y})$. Hence lift is a ratio of observed confidence to the confidence that is expected under the assumption of independence.

Another view angle on lift is provided by the fact that

$$\mathrm{lift}(X \rightharpoonup Y) = \mathrm{lift}(Y \rightharpoonup X) = \frac{\mathrm{supp}(X \rightharpoonup Y)}{\mathrm{supp}(X) \cdot \mathrm{supp}(Y)}.$$

If X and Y are stochastically independent, $\mathrm{P}(\mathsf{X} \wedge \mathsf{Y}) = \mathrm{P}(\mathsf{X}) \cdot \mathrm{P}(\mathsf{Y})$. Lift is therefore a ratio $\frac{\mathrm{P}(\mathsf{X} \wedge \mathsf{Y})}{\mathrm{P}(\mathsf{X}) \cdot \mathrm{P}(\mathsf{Y})}$ of the observed probability and the probability expected under the assumption of independence of X and Y, or, in other words, lift is a ratio of observed and expected support.

Generally, everyone is interested in association rules with non-zero support. For that rules, lift is always greater than 0. If X and Y are independent, $\mathrm{lift}(X \rightharpoonup Y)$ equals 1. Values greater than 1 indicate positive dependency, values lower than 1 indicate negative dependency.

2.2 Fuzzy Association Rules

For fuzzy association rules, domain of each fuzzy attribute $a \in \mathcal{A}$ is not binary (or "crisp") $\{0, 1\}$, but graded (or "fuzzy"), i.e. interval $[0, 1]$. That is, for each $a \in \mathcal{A}$ and $o \in \mathcal{O}$, $a(o) \in [0, 1]$. For a subset $X \subseteq \mathcal{A}$ of fuzzy attributes, we define a new predicate of a logical conjunction (similarly to binary case (1)) by using a t-norm $\otimes$:

$$X(o_j) \;:=\; \bigotimes_{a \in X} a(o_j). \tag{6}$$

T-norm $\otimes$ is a generalized logical conjunction, i.e. a function $[0, 1] \times [0, 1] \to [0, 1]$ which is associative, commutative, monotone increasing (in both places) and which satisfies the boundary conditions $\alpha \otimes 0 = 0$ and $\alpha \otimes 1 = \alpha$ for each $\alpha \in [0, 1]$. Some well-known examples of t-norms are:

- product t-norm: $\otimes_{\mathrm{prod}}(\alpha, \beta) = \alpha\beta$;
- minimum t-norm: $\otimes_{\mathrm{min}}(\alpha, \beta) = \min(\alpha, \beta)$;
- Łukasiewicz t-norm: $\otimes_{\text{Łuk}}(\alpha, \beta) = \max(0, \alpha + \beta - 1)$.

Let $a \in \mathcal{A}$, $o \in \mathcal{O}$, $n = |\mathcal{O}|$, $n > 0$, and $X, Y \subset \mathcal{A}$, $X \neq \emptyset$, $Y \neq \emptyset$, $X \cap Y = \emptyset$. Several quality measures may be defined as follows:

$$\mathrm{fsupp}(X) := \frac{\sum_{o \in \mathcal{O}} X(o)}{n}, \tag{7}$$

$$\mathrm{fsupp}(X \rightharpoonup Y) := \mathrm{fsupp}(X \cup Y), \tag{8}$$

$$\mathrm{fconf}(X \rightharpoonup Y) := \frac{\mathrm{fsupp}(X \rightharpoonup Y)}{\mathrm{fsupp}(X)}. \tag{9}$$

3 Lift on Fuzzy Association Rules

A naive approach for introducing lift to the fuzzy association rules framework is to use simply the definition of lift (5) for binary rules and replace binary support (2, 3) and confidence (4) with their fuzzy alternatives (7, 8, 9) as e.g. in [5,15]. Unfortunately, that approach works only if product t-norm is in use (see section 3.1). If using minimum or Łukasiewicz t-norms, this may lead to errors (see sections 3.2 and 3.3).

In this section, a proper definition of lift for fuzzy association rules is presented. Later, some features of that definition are studied.

As noted in Section 2.1, lift can be understood as a ratio of the observed support $\mathrm{fsupp}(X \rightharpoonup Y)$ to the expected support $\mathrm{E}\left[\mathrm{fsupp}(X \rightharpoonup Y)\right]$. The observed support is simply (8), but, given sets X and Y of fuzzy attributes, what support is expected if X and Y are independent? Moreover, what does independency of fuzzy attributes mean?

For the sake of simplicity, let us assume X and Y be sets containing a single fuzzy attribute, $|X| = |Y| = 1$. For more complex cases, a new attribute can be created from the set of fuzzy attributes by using (6).

If objects $o \in \mathcal{O}$ are selected randomly, one can treat the membership values, $X(o)$ and $Y(o)$, as random variables X and Y, and treat the independence of fuzzy attributes as stochastic independence of random variables X and Y. Two random variables X, Y are stochastically independent, if the combined random variable $\big(\mathsf{X}, \mathsf{Y}\big)$ has a joint probability density

$$f_{\mathsf{X},\mathsf{Y}}(x, y) = f_{\mathsf{X}}(x) f_{\mathsf{Y}}(y). \tag{10}$$

If X and Y are two independent random variables from interval $[0, 1]$ then

$$\sigma(x, y) := \frac{x \otimes y}{n} \tag{11}$$

is a random variable with probability density function $f_\sigma(x, y) = f_{\mathsf{X},\mathsf{Y}}(x, y)$.

Generally, *expected value* $\mathrm{E}\left[\mathsf{Z}\right]$ of a random variable Z is a weighted average of all possible values. More formally, $\mathrm{E}\left[\mathsf{Z}\right] = \int_{-\infty}^{\infty} z f_{\mathsf{Z}}(z) \mathrm{d}z$, where f_{Z} is a probability density function of random variable Z.

Similarly, an expected value $\mathrm{E}\left[\sigma(x, y)\right]$ is a weighted average of all possible (x, y) pairs, namely

$$\mathrm{E}\left[\sigma(x, y)\right] = \int_0^1 \int_0^1 \sigma(x, y) f_\sigma(x, y)\ \mathrm{d}x \mathrm{d}y. \tag{12}$$

In reality, $f_\sigma(x, y)$ is unknown, but we can estimate its values from data (i.e. from objects $\mathcal{O}$ and their fuzzy attributes $\mathcal{A}$) by using the assumption of independence (10):

$$f_\sigma(x, y) = f_{\mathsf{X}}(x) f_{\mathsf{Y}}(y) \approx \frac{\mathrm{count}_X(x)}{n} \cdot \frac{\mathrm{count}_Y(y)}{n}, \tag{13}$$

where $\mathrm{count}_A(a)$ is the number of objects from $\mathcal{O}$ that belong to A with degree a, i.e. $\mathrm{count}_A(a) = \big|\{o \in \mathcal{O} | A(o) = a\}\big|$.

Assuming $x \in \{X(o)|o \in \mathcal{O}\}$ and $y \in \{Y(o)|o \in \mathcal{O}\}$, we obtain

$$\mathrm{E}\left[\sigma(x,y)\right] \approx \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{X(o_i) \otimes Y(o_j)}{n^3}$$

by inserting (11), (13) into (12). Since X and Y are independent, $\mathrm{E}\left[\mathrm{fsupp}(X \rightharpoonup Y)\right] = n \cdot \mathrm{E}\left[\sigma(x,y)\right]$ and hence

$$\mathrm{E}\left[\mathrm{fsupp}(X \rightharpoonup Y)\right] \approx \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{X(o_i) \otimes Y(o_j)}{n^2}. \tag{14}$$

Definition 1. *Let $\otimes$ be a t-norm, X, Y be sets of fuzzy attributes such that* $\mathrm{fsupp}(X) > 0$ *and* $n > 0$. *Then the* expected fuzzy support $\mathrm{fs\widehat{upp}}(X \rightharpoonup Y)$ *and the* expected fuzzy confidence $\mathrm{f\widehat{con}f}(X \rightharpoonup Y)$ *are defined as follows:*

$$\mathrm{fs\widehat{upp}}(X \rightharpoonup Y) := \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{X(o_i) \otimes Y(o_j)}{n^2},$$

$$\mathrm{f\widehat{con}f}(X \rightharpoonup Y) := \frac{\mathrm{fs\widehat{upp}}(X \rightharpoonup Y)}{\mathrm{fsupp}(X)}.$$

Theorem 1. *Let X, Y be sets of fuzzy attributes. Then:*

1. *if* $\mathrm{fsupp}(X \rightharpoonup Y) > 0$ *then* $\mathrm{fs\widehat{upp}}(X \rightharpoonup Y) > 0$,
2. $\mathrm{fs\widehat{upp}}(X \rightharpoonup Y) \leq \min(\mathrm{fsupp}(X), \mathrm{fsupp}(Y))$.

Proof. 1) If $\mathrm{fsupp}(X \rightharpoonup Y) > 0$ then $\sum_{i=1}^{n} X(o_i) \otimes Y(o_i) > 0$ and hence also $\sum_{i=1}^{n}\sum_{j=1}^{n} X(o_i) \otimes Y(o_j) > 0$. Therefore $\mathrm{fs\widehat{upp}}(X \rightharpoonup Y) > 0$.

2) For any t-norm $\otimes$, $X(o_i) \otimes Y(o_j) \leq Y(o_j)$. Therefore:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{X(o_i) \otimes Y(o_j)}{n^2} \leq \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{Y(o_j)}{n^2} = \mathrm{fsupp}(Y),$$

and similarly for $\mathrm{fsupp}(X)$. Hence $\mathrm{fs\widehat{upp}}(X \rightharpoonup Y) \leq \min(\mathrm{fsupp}(X), \mathrm{fsupp}(Y))$.

Definition 2. *Let X, Y be sets of fuzzy attributes such that* $\mathrm{fsupp}(X \rightharpoonup Y) > 0$ *and* $n > 0$. *Then* fuzzy lift *of rule $X \rightharpoonup Y$ is defined as follows:*

$$\mathrm{flift}(X \rightharpoonup Y) := \frac{\mathrm{fsupp}(X \rightharpoonup Y)}{\mathrm{fs\widehat{upp}}(X \rightharpoonup Y)}.$$

We assume $\otimes$ be an arbitrary (but fixed) t-norm. Where it is important to explicitly denote the concrete used t-norm $\otimes$, we put $\otimes$ in subscript and write e.g. $\mathrm{flift}_{\otimes}(X \rightharpoonup Y)$ instead of $\mathrm{flift}(X \rightharpoonup Y)$.

In accordance with the discussion above, if the (sets of) fuzzy attributes X and Y are independent, the value of $\mathrm{flift}(X \rightharpoonup Y)$ is close to 1. For $\mathrm{flift}(X \rightharpoonup Y) > 1$ (resp. < 1), there is higher (resp. lower) occurence of $X \cup Y$ than expected, therefore $\mathrm{flift}(X \rightharpoonup Y) > 1$ (resp. < 1) indicates positive (resp. negative) dependency among X and Y.

Definition 3. *We call sets of fuzzy attributes X and Y to be:*

1. positively dependent *if* $\mathrm{flift}(X \rightharpoonup Y) > 1$*;*
2. negatively dependent *if* $\mathrm{flift}(X \rightharpoonup Y) < 1$*.*

The order of X and Y in the previous definition is not important, because, as can be seen in the subsequent theorem, $\mathrm{flift}(X \rightharpoonup Y) = \mathrm{flift}(Y \rightharpoonup X)$.

Theorem 2. *Let X, Y be sets of fuzzy attributes. Then:*

1. $\mathrm{flift}(X \rightharpoonup Y) = \mathrm{flift}(Y \rightharpoonup X)$*,*
2. $\mathrm{flift}(X \rightharpoonup Y) = \frac{\mathrm{fconf}(X \rightharpoonup Y)}{\mathrm{f}\widehat{\mathrm{conf}}(X \rightharpoonup Y)}$*,*
3. $0 \leq \mathrm{flift}_{\otimes}(X \rightharpoonup Y) \leq n$*,*
4. *if* $\mathrm{fsupp}(X \rightharpoonup Y) > 0$ *then* $\mathrm{flift}(X \rightharpoonup Y) > 0$*.*

Proof. 1) and 2) directly follow from the definitions and from the fact that t-norms are commutative.

3) Since the membership degrees are defined on interval $[0, 1]$, their sums cannot be negative either. Hence $\mathrm{flift}(X \rightharpoonup Y) \geq 0$. Next, assume to the contrary that $\mathrm{flift}(X \rightharpoonup Y) > n$. Then $\sum_{i=1}^{n} X(o_i) \otimes Y(o_i) > \sum_{i=1}^{n} \sum_{j=1}^{n} X(o_i) \otimes Y(o_j)$, which is a contradiction.

4) If $\mathrm{fsupp}(X \rightharpoonup Y) > 0$ then from Theorem 1 we know that also $\mathrm{fs}\widehat{\mathrm{upp}}(X \rightharpoonup Y) > 0$. Therefore $\mathrm{flift}(X \rightharpoonup Y)$ exists and is greater than 0.

Theorem 2 shows fuzzy lift's properties that are analogous to those of crisp lift that were discussed in section 2.1.

3.1 Fuzzy Lift with Product T-norm

In this sub-section, properties of fuzzy lift are studied for the special case of $\otimes := \otimes_{\mathrm{prod}}$, i.e. for the product t-norm being used.

Theorem 3. *Let X, Y be sets of fuzzy attributes such that* $\mathrm{fsupp}(X \rightharpoonup Y) > 0$ *and $n > 0$. Then:*

$$\mathrm{flift}_{\otimes_{prod}}(X \rightharpoonup Y) = \frac{\mathrm{fsupp}(X \rightharpoonup Y)}{\mathrm{fsupp}(X) \cdot \mathrm{fsupp}(Y)}.$$

Proof. If $\mathrm{fsupp}(X \rightharpoonup Y) > 0$, then evidently $\mathrm{fsupp}(X) > 0$ and $\mathrm{fsupp}(Y) > 0$. Assume $\otimes := \otimes_{\mathrm{prod}}$, then: $\mathrm{fs}\widehat{\mathrm{upp}}(X \rightharpoonup Y) = \frac{\sum_{i=1}^{n} X(o_i) \sum_{i=1}^{n} Y(o_i)}{n^2} = \mathrm{fsupp}(X) \cdot \mathrm{fsupp}(Y)$.

Note that Theorem 3 holds for product t-norm only. For other t-norms such as minimum or Łukasiewicz t-norm, the Definition 2 of fuzzy lift must not be over-simplified that way.

3.2 Fuzzy Lift with Minimum T-norm

If minimum t-norm is used (i.e. $\otimes := \otimes_{\min}$), the following theorem holds.

Theorem 4. *Let X, Y be sets of fuzzy attributes such that* $\mathrm{fsupp}(X \rightharpoonup Y) > 0$ *and $n > 0$. Then:*

$$\frac{\mathrm{fsupp}(X \rightharpoonup Y)}{\min(s_X, s_Y)} \leq \mathrm{flift}_{\otimes_{min}}(X \rightharpoonup Y) \leq \frac{\mathrm{fsupp}(X \rightharpoonup Y)}{s_X \cdot s_Y},$$

where $s_X = \mathrm{fsupp}(X)$ and $s_Y = \mathrm{fsupp}(Y)$.

Proof. The first inequality directly follows from theorem 1. The second inequality follows from $\otimes_{\min}(x, y) \geq x \cdot y$: $\widehat{\mathrm{fsupp}}(X \rightharpoonup Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{X(o_i) \otimes_{\min} Y(o_j)}{n^2} \geq \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{X(o_i) \cdot Y(o_j)}{n^2} = s_X \cdot s_Y$. Note also, if $\mathrm{fsupp}(X \rightharpoonup Y) > 0$, then evidently $\mathrm{fsupp}(X) > 0$ and $\mathrm{fsupp}(Y) > 0$.

3.3 Fuzzy Lift with Łukasiewicz T-norm

Finally, fuzzy lift's properties are studied if defined with the Łukasiewicz $\otimes_{\text{Łuk}}$ t-norm.

Lemma 1. *Let X, Y be sets of fuzzy attributes, $n > 0$. Then:*

$$\max_{i \in \{1,\ldots,n\}} (X(o_i)) + \max_{i \in \{1,\ldots,n\}} (Y(o_i)) \leq 1 \quad \textit{iff} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} X(o_i) \otimes_{\text{Łuk}} Y(o_j) = 0.$$

Proof. Let $\max_{\forall i}(X(o_i)) + \max_{\forall i}(Y(o_i)) \leq 1$, then $\forall i, j \in \{1, 2, \ldots, n\}$, $X(o_i) \otimes_{\text{Łuk}} Y(o_j) = 0$ and hence $\sum_{i=1}^{n} \sum_{j=1}^{n} X(o_i) \otimes_{\text{Łuk}} Y(o_j) = 0$.

Let now $\sum_{i=1}^{n} \sum_{j=1}^{n} X(o_i) \otimes_{\text{Łuk}} Y(o_j) = 0$ and let us take such i, j that $X(o_i)$ is maximum among X and $Y(o_j)$ is maximum among Y. Then also $X(o_i) \otimes_{\text{Łuk}} Y(o_j) = 0$, hence $\max_{\forall i}(X(o_i)) + \max_{\forall i}(Y(o_i)) \leq 1$.

Theorem 5. *Let X, Y be sets of fuzzy attributes with* $\mathrm{fsupp}(X) = s_X$ *and* $\mathrm{fsupp}(Y) = s_Y$ *and let* $\mathrm{fsupp}(X \rightharpoonup Y) > 0$*, $n > 0$. Then:*

$$\frac{\mathrm{fsupp}(X \rightharpoonup Y)}{s_X \cdot s_Y} \leq \mathrm{flift}_{\otimes_{\text{Łuk}}}(X \rightharpoonup Y) \leq \frac{\mathrm{fsupp}(X \rightharpoonup Y)}{s_X \otimes_{\text{Łuk}} s_Y}.$$

Proof. The first inequality follows from $\otimes_{\text{Łuk}}(x, y) \leq x \cdot y$, because then we have $\widehat{\mathrm{fsupp}}(X \rightharpoonup Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{X(o_i) \otimes_{\text{Łuk}} Y(o_j)}{n^2} \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{X(o_i) \cdot Y(o_j)}{n^2} = s_X \cdot s_Y$.

To prove the second inequality, it suffices to prove

$$\widehat{\mathrm{fsupp}}(X \rightharpoonup Y) \geq \max(0, s_X + s_Y - 1), \tag{15}$$

which is obvious for $s_X + s_Y \leq 1$. Let us therefore assume $s_X + s_Y > 1$, then (15) can be rewritten as: $\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{X(o_i) \otimes_{\text{Łuk}} Y(o_j)}{n^2} \geq s_X + s_Y - 1$. The double sum $\sum_{i=1}^{n} \sum_{j=1}^{n} X(o_i) \otimes_{\text{Łuk}} Y(o_j)$ equals $\sum_{i=1}^{n} \sum_{j=1}^{n} \big(X(o_i) + Y(o_j)\big) - \sum_{\forall k}(t_k) -$

Table 1. Examples of How Dependency Orientation Can Change If Using Different T-norms

X			Y			$\text{flift}_{\otimes}(X \rightharpoonup Y)$			dependency of X and Y		
o_1	o_2	o_3	o_1	o_2	o_3	$\otimes_{\text{min}}$	$\otimes_{\text{prod}}$	$\otimes_{\text{Łuk}}$	$\otimes_{\text{min}}$	$\otimes_{\text{prod}}$	$\otimes_{\text{Łuk}}$
0.99	0.04	0.22	0.25	0.27	0.01	0.84	1.18	–	neg.	pos.	–
0.17	0.04	0.00	0.26	0.08	0.93	1.17	0.56	–	pos.	neg.	–
0.11	0.36	0.44	0.61	0.92	0.05	–	0.88	1.18	–	neg.	pos.
0.12	0.25	0.19	0.12	0.52	0.83	–	1.10	0.55	–	pos.	neg.
0.78	0.03	0.97	0.02	0.25	0.31	0.86	–	1.33	neg.	–	pos.
0.17	0.09	0.54	0.96	0.25	0.54	1.13	–	0.81	pos.	–	neg.

$|r|$, where t is a sequence of numbers $\big(X(o_i) + Y(o_j) | X(o_i) + Y(o_j) < 1\big)$ and $r = \{(i,j) | X(o_i) + Y(o_j) >= 1\}$, for $i, j \in \{1, 2, \ldots, n\}$. Since $|t| + |r| = n^2$ and each $t_k < 1$, we can immediately see that $\sum_{i=1}^{n} \sum_{j=1}^{n} X(o_i) \otimes_{\text{Łuk}} Y(o_j) \geq \sum_{i=1}^{n} \sum_{j=1}^{n} \big(X(o_i) + Y(o_j)\big) - n^2 = n^2 s_X + n^2 s_Y - n^2 = n^2(s_X + s_Y - 1)$, hence (15) holds.

3.4 Comparison of T-norms w.r.t. Fuzzy Lift

Observation: Fuzzy lift does not preserve order, if switching t-norms. Even worse, positive dependency may be changed to negative (and vice versa), only by selecting a different t-norm. (Positive/negative dependency is understood as in definition 3).

Please consider Table 1: each row is an example of the X and Y attributes' membership degrees for objects o_1, o_2, o_3. Columns 7–9 are fuzzy lifts computed for the minimum $\otimes_{\text{min}}$, product $\otimes_{\text{prod}}$ or Łukasiewicz $\otimes_{\text{Łuk}}$ t-norm being used, respectively. Unimportant values are omitted with "–".

As you can see, there can be found such fuzzy sets, for whose the used t-norm derermines whether the resulting fuzzy lift shows positive or negative dependency. For instance, data from the first row result in $\text{flift}_{\otimes_{\text{min}}}(X \rightharpoonup Y)$ indicating negative dependency (< 1), whereas $\text{flift}_{\otimes_{\text{prod}}}(X \rightharpoonup Y)$ evaluated on the same data indicates positive dependency (> 1). Second row shows the opposite situation, and so on.

4 Fast Computation of Fuzzy Lift

Besides combinatorial explosion caused by rule generation, a time-effectivity bottleneck of association rules mining algorithms (both crisp and fuzzy) is computation of rule support [6], because a relatively slow scan of a source dataset has to be performed for each itemset [2], [1]. Once computed, support is used in search tree pruning conditions and for computations of other measures such as crisp (4) and fuzzy confidence (9). Also crisp lift (5) and fuzzy lift with the product t-norm (see Theorem 3) can be computed directly from support. Unfortunately, it seems not to be the case for fuzzy lift with minimum or Łukasiewicz t-norms.

Algorithm 1. Fast computation of fuzzy lift with minimum or Łukasiewicz t-norms.

Variables that are assumed to exist:

n – number of rows in dataset ($n := |\mathcal{O}|$);
s_a – fuzzy support of a (based on $\otimes_{\min}$ for $\textsc{flift}_{min}$, resp. $\otimes_{\text{Łuk}}$ for $\textsc{flift}_{Łuk}$); we assume $s_a > 0$ for each $a \in \mathcal{A}$;
$t_a[i]$ – membership degrees $a(o)$ (for $o \in \mathcal{O}$) sorted in ascending order (for $i \in \{1, 2, \ldots, n\}$ and $t_a[0] := 0$); only needed for any $a \in \mathcal{A}$ that may appear in consequent;
c_a – array of cummulative sums of t_a, i.e. $c_a[0] := t_a[0]$, $c_a[i] := t_a[i] + c_a[i-1]$ (for $i \in \{1, 2, \ldots, n\}$).

```
function FLIFT$_{min}(x, y)$
    $r \leftarrow 0$
    for $i \in \{1, 2, \ldots, n\}$ do
        $l_i \leftarrow$ index of $x[i]$ in $t_y$ (or of largest smaller value) found with binary search
        $r \leftarrow r + x[i] \cdot (n - l_i) + c_y[l_i]$
    end for
    return $n^2 s_{xy}/r$
end function

function FLIFT$_{Łuk}(x, y)$
    $r \leftarrow 0$
    for $i \in \{1, 2, \ldots, n\}$ do
        $l_i \leftarrow$ index of $(1 - x[i])$ in $t_y$ (or of largest smaller value) found with binary search
        $r \leftarrow r + x[i] \cdot (n - l_i) + c_y[n] - c_y[l_i] - n + l_i$
    end for
    return $n^2 s_{xy}/r$
end function
```

A naive approach for computation of fuzzy lift with $\otimes_{\min}$ or $\otimes_{\text{Łuk}}$, for single rule $X \rightharpoonup Y$, uses Definition 2 and therefore leads to the $O(n^2)$ time complexity algorithm, because of the nested sums in the definition of expected fuzzy support $\widehat{\text{fsupp}}(A \rightharpoonup B)$. Here n is the number of rows in a dataset, i.e. n is typically a very large number.

In this section, an algorithm for fuzzy lift evaluation is presented that has $O(n \log n)$ time complexity.

The algorithm assumes all attributes $a \in \mathcal{A}$ that may appear in rule's consequent to be preprocessed as follows:

1. first, all membership degrees of a are sorted in ascending order and stored into array t_a indexed from 1; moreover, $t_a[0] := 0$;

2. next, an array c_a of cummulative sums of values in t_a is computed: $c_a[0] := t_a[0]$, $c_a[i] := t_a[i] + c_a[i-1]$ (for $i \in \{1, 2, \ldots, n\}$).

These preprocessing steps can be computed only once, at the beginning of the rule mining process.

It is also expected the underlying association rules searching algorithm has already computed fuzzy supports $s_{XY} := \text{fsupp}(X \rightharpoonup Y)$ in order to perform pruning and computations of other interest measures such as fuzzy confidence (9). The algorithm is valid only for $s_{XY} > 0$, since for zero support, the fuzzy lift is undefined.

For $\otimes_{\min}$, the computation of fuzzy lift (FLIFT$_{min}$ in Algorithm 1) runs as follows. Firstly, r is computed in steps 2 to 6 so that

$$r = \sum_{i=1}^{n} \left(x[i](n - l_i) + \sum_{k=0}^{l_i} t_y[k] \right), \tag{16}$$

where l_i is such index that $t_y[l_i] \leq x[i] < t_y[l_i+1]$. Because we have set $t_y[0] := 0$, formula (16) can be rewritten as $\sum_{i=1}^{n} \sum_{j=1}^{n} \min(x[i], y[i])$, hence we are convinced that FLIFT$_{min}$ really returns $\text{flift}_{\otimes_{\min}}(X \rightharpoonup Y)$.

Regarding FLIFT$_{\text{Ł}uk}$, the value of r computed in steps 10 to 14 equals to

$$\sum_{i=1}^{n} \left(x[i](n - l_i) + \sum_{j=1}^{n} t_y[j] - \sum_{k=0}^{l_i} t_y[k] - n + l_i \right), \tag{17}$$

where l_i is now such index that $t_y[l_i] \leq (1 - x[i]) < t_y[l_i + 1]$. Formula (17) is equivalent to $\sum_{i=1}^{n} \Big(nx[i] - l_i x[i] + \sum_{j=1}^{n} y[j] - \sum_{k=0}^{l_i} t_y[k] - n + l_i \Big)$, which can be rewritten to $\sum_{i=1}^{n} \Big(\sum_{j=1}^{n} \big(x[i] + y[j] - 1 \big) - \sum_{k=0}^{l_i} t_y[k] - l_i x[i] + l_i \Big)$, which in turn equals to $\sum_{i=1}^{n} \sum_{j=1}^{n} \max(0, x[i] + y[i] - 1)$, and it proves that FLIFT$_{\text{Ł}uk}$ returns $\text{flift}_{\otimes_{\text{Łuk}}}(X \rightharpoonup Y)$.

Let us now analyze time and space complexity. Step 4 (resp. 12) performs binary search in an ordered array, which is known to have $O(\log n)$ time complexity. Together with the for-loop in step 3 (resp. 11) it gives the overall time complexity of $O(n \log n)$.

For FLIFT$_{min}$ (resp. FLIFT$_{\text{Ł}uk}$) to work properly, we need to have pre-computed the arrays t_a and s_a. It is known that the time complexity of a sort algorithm is $O(n \log n)$ (for t_a), whereas computing cummulative sums s_a can be done in $O(n)$. Moreover, that steps need to be done only once at the beginning of the association rules mining algorithm and then shared accross multiple calls of FLIFT$_{min}$ (resp. FLIFT$_{\text{Ł}uk}$). Taking all of that into account, the time complexity still remains $O(n \log n)$.

In addition, two arrays (per consequent fuzzy attribute a), t_a and c_a, need to be stored in memory. Therefore, the space complexity of fuzzy lift computation is in $O(2m'n)$, where m' is the number of fuzzy attributes that may appear in any rule's consequent.

5 Conclusion

Lift is a ratio of observed support (resp. confidence) to the support (resp. confidence) that is expected under the assumption of independence. In this paper, a correct definition of fuzzy lift was provided. It should be stressed here that there already exist some research papers that use incorrect definition of fuzzy lift (e.g. [15]).

Besides definition, some interesting mathematical properties of fuzzy lift were studied and the values of lift were compared if computed with different t-norms.

Fuzzy lift has equivalent definition to the "crisp" lift (i.e. lift on binary data) if the t-norm being used is product $\otimes_{\mathrm{prod}}$. For Łukasiewicz $\otimes_{\mathrm{Luk}}$ and minimum $\otimes_{\min}$ t-norms, a more complicated computation takes place. Therefore, an algorithm was developed in Section 4 for fast evaluation of fuzzy lift. It has been also proven that the algorithm's time complexity is in $O(n \log n)$, for n being the number of objects in dataset, while the space complexity is linear with respect to the number of fuzzy attributes that may appear in rule consequents. (An equivalent naive algorithm's time complexity is $O(n^2)$; space complexity is linear.)

A future research will address improvements of associtions rules search algorithms by introducing heuristics based on boundary conditions provided by Theorems 4 and 5. The idea is to store rules with best lift only and to not to traverse through fuzzy attribute combinations that do not have a potential to provide lift that is good-enough to the user. Also other interest measures may be studied and their applicability on fuzzy rules considered.

Acknowledgment. This work was supported by the European Regional Development Fund in the project of IT4Innovations Centre of Excellence (CZ.1.05/1.1.00/02.0070, VP6). The author would like to express gratitude also to Viktor Pavliska (IRAFM) for his constructive ideas and comments.

References

1. Agrawal, R.: Fast discovery of association rules. In: Advances in Knowledge Discovery and Data Mining, pp. 307–328. AAAI Press / MIT Press (1996)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining associations between sets of items in massive databases. In: ACM SIGMOD 1993 Int. Conference on Management of Data, Washington, D.C, pp. 207–216 (1993)
3. Berrado, A., Runger, G.C.: Using metarules to organize and group discovered association rules. Data Min. Knowl. Discov. 14(3), 409–431 (2007)
4. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA, pp. 255–264 (May 1997)
5. Buczak, A.L., Gifford, C.M.: Fuzzy association rule mining for community crime pattern discovery. In: ACM SIGKDD Workshop on Intelligence and Security Informatics, ISI-KDD 2010, pp. 2:1–2:10. ACM, New York (May 2010)

6. Burda, M.: Fast evaluation of t-norms for fuzzy association rules mining. In: 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI 2013), pp. 465–470. IEEE, Budapest (2013)
7. Chan, K.C., Au, W.H.: Mining fuzzy association rules (1997)
8. Hahsler, M., Hornik, K.: New probabilistic interest measures for association rules. Intell. Data Anal. 11(5), 437–455 (2007)
9. Hájek, P., Havel, I., Chytil, M.: The GUHA method of automatic hypotheses determination. Computing 1, 293–308 (1966)
10. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Min. Knowl. Discov. 8(1), 53–87 (2004)
11. Kalia, H., Dehuri, S., Ghosh, A.: A survey on fuzzy association rule mining. International Journal of Data Warehousing and Mining (IJDWM) 9(1), 1–27 (2013)
12. Lallich, S., Teytaud, O., Prudhomme, E.: Association rule interestingness: Measure and statistical validation. In: Guillet, F., Hamilton, H.J. (eds.) Quality Measures in Data Mining. SCI, vol. 43, pp. 251–275. Springer, Heidelberg (2007)
13. McNicholas, P.D., Murphy, T.B., O'Regan, M.: Standardising the lift of an association rule. Comput. Stat. Data Anal. 52(10), 4712–4721 (2008)
14. Novák, V., Perfilieva, I., Dvořák, A., Chen, G., Wei, Q., Yan, P.: Mining pure linguistic associations from numerical data. Int. J. Approx. Reasoning 48(1), 4–22 (2008)
15. Pancho, D.P., Alonso, J.M., Alcalá-Fdez, J., Magdalena, L.: Interpretability analysis of fuzzy association rules supported by fingrams. In: EUSFLAT Conf. (2013)
16. Pei, B., Zhao, S., Chen, H., Zhou, X., Chen, D.: Farp: Mining fuzzy association rules from a probabilistic quantitative database. Information Sciences 237, 242–260 (2013)
17. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. SIGMOD Rec. 25(2), 1–12 (1996)
18. Webb, G.I.: Discovering significant patterns. Mach. Learn. 68(1), 1–33 (2007)

Fuzzy Rule-Based Ensemble for Time Series Prediction: Progresses with Associations Mining

Michal Burda*, Martin Štěpnička, and Lenka Štěpničková

Institute for Research and Applications of Fuzzy Modeling, University of Ostrava,
Centre of Excellence IT4Innovations, 30. dubna 22, Ostrava, Czech Republic
{Michal.Burda,Martin.Stepnicka,Lenka.Stepnickova}@osu.cz

Abstract. As there are many various methods for time series prediction developed but none of them generally outperforms all the others, there always exists a danger of choosing a method that is inappropriate for a given time series. To overcome such a problem, distinct ensemble techniques, that combine more individual forecasts, are being proposed. In this contribution, we employ the so called fuzzy rule-based ensemble. This method is constructed as a linear combination of a small number of forecasting methods where the weights of the combination are determined by fuzzy rule bases based on time series features such as trend, seasonality, or stationarity. For identification of fuzzy rule base, we use linguistic association mining. An exhaustive experimental justification is provided.

Keywords: Fuzzy rule-based ensemble, time series, fuzzy rules, ensemble, perception-based logical deduction, linguistic associations mining.

1 Introduction

A time series is given as a finite sequence $y_1, y_2, \ldots, y_T$ of real numbers and the task is to predict future values $y_{T+1}, y_{T+2}, \ldots, y_{T+h}$ where h denotes so called *forecasting horizon*. There are many different methods for this task that are nowadays widely used in practice. Unfortunately, there is no single forecasting method that generally outperforms any other. Thus, there is a danger of choosing a method which is inappropriate for a given time series. Note that even searching for methods, that outperform any other for narrower specific subsets of time series, has not been successful yet, see e.g. [2], where the authors stated: *"Although forecasting expertise can be found in the literature, these sources often fail to adequately describe conditions under which a method is expected to be successful"*.

In order to eliminate the risk of choosing an inappropriate method, distinct *ensemble techniques* (*ensembles* in short) have been designed and successfully

* This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). Furthermore, we acknowledge partial support of projects KONTAKT II - LH12229 of MŠMT ČR.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_31

applied. The main idea of ensembles consists in an appropriate combination of more forecasting methods. Typically, an ensemble technique is constructed as a linear combination of individual ones. It can be described as follows. Let us assume that we are given a set of M individual methods and let for a given times series $y_1, y_2, \ldots, y_T$ and a given forecasting horizon h, the j-th individual method provides us with the following prediction:

$$\hat{y}^{(j)}_{T+1}, \hat{y}^{(j)}_{T+2}, \ldots, \hat{y}^{(j)}_{T+h}, \quad j = 1, \ldots, M.$$

Then the ensemble forecast is given by the following formula:

$$\hat{y}_{T+i} = \frac{1}{\sum_{j=1}^{M} w_j} \cdot \sum_{j=1}^{M} w_j \cdot \hat{y}^{(j)}_{T+i}, \quad i = 1, \ldots, h,$$

where $w_j \in \mathbb{R}$ is a weight of the j-th individual method. These weights are usually normalized, that is, $\sum_{j=1}^{M} w_j = 1$.

Let us recall that it was perhaps Bates and Granger [4] who firstly showed significant gains in accuracy through combinations. Another early work by Newbold and Granger [19] combined various time series forecasts and compared the combination against the performance of the individual methods. They showed that for set of forecasts, a linear combination of these forecasts achieved a forecast error variance smaller than the individual forecasts. They found that the better combining procedures did produce an overall forecast superior to individual forecasts on the majority of tested time series.

How to combine methods, i.e., how to determine appropriate weights, is still a relatively open question. For instance, Makridakis et al. [16] showed that taking a simple average alias the so called *"equal-weights combining"* [6], is a benchmark that is hard to beat and finding appropriate non-equal weights leads rather to a damage of the averaging idea that causes the improvements in robustness.

Although the equal-weights ensemble performs as accurately as mentioned above, there are works that promisingly show the potential of more sophisticated approaches. We recall [15] that described an approach using meta-learning for time series forecasting based on the features of time series such as: standard deviation, skewness, etc. Given time series were clustered and individual methods were ranked according to their performance on each cluster and then three best methods for each cluster were selected. For a given new time series, the closest cluster was determined and the given three best methods were combined.

This approach was one of our main motivations because it demonstrates that there exists a dependence between time series features and a performance of a forecasting method. The second major motivation stems from the so called *Rule-Based Forecasting* (RBF) developed by Collopy and Armstrong [2,6]. It is an expert system that uses domain knowledge to combine forecasts from various forecasting methods. Using IF-THEN rules, RBF determines what weights to give to the forecasts.

We follow the main ideas of rule-based forecasting [2] and of using time series features [15] to obtain an interpretable and understandable ensemble model.

2 Fuzzy Rule-Based Ensemble

As mentioned above, RBF uses the rules to determine weights [2]. However, only few of these rules are directly used to set up weights. Most of them set up a rather specific model parameters, e.g. the smoothing factors of the Brown's exponential smoothing with trend. Moreover, in antecedents, the rules very often use properties that are not crisp but rather vague, e.g. expressions such as: "trend has been changing; unstable recent trend" etc., see [6]. For such cases, using crisp rules seems to be less natural than using fuzzy rules. Similarly, the use of crisp consequents such as: "add 10% to the weight; subtract 0.4 from beta" etc., seems to be less intuitive than using vague expressions that are typical for fuzzy rules.

2.1 General Structure of the Model

Therefore, our goal was to propose a method that uses fuzzy rules instead of crisp rules in order to capture the omnipresent vagueness in the expressions; to use only quantitative features (no domain knowledge) in the antecedent variables which enable to fully automatize the method; to use only individual forecasting method weights as the consequent variables. The result of such motivated investigation is the Fuzzy Rule-Based Ensemble (FRBE) [23,26].

The FRBE method uses a single *linguistic description*, i.e. fuzzy rule base with *evaluative linguistic expressions* [21], to determine a weight of each forecasting method based on fuzzy/linguistic rules, such as:

"**IF** *Strength of Seasonality is Small* **AND** *Coefficient of Variation is Roughly Small* **THEN** *Weight of the j-th method is Big*".

After an appropriate inference method is applied (see Section 2.2), a defuzzification method is employed and thus, a crisp result (weight of a particular method) is determined. All such weights are then used to determine the final combined output as given by (1).

2.2 Components of the Model

In order to estimate (set up) a particular value of the weight of each forecasting method with help of the fuzzy rules, an appropriate fuzzy inference mechanism has to be employed. As mentioned above, the FRBE method employs linguistic descriptions, i.e. fuzzy rule bases with so called evaluative linguistic expressions. These are expressions of natural language that are based on the expressions of the basic trichotomy `Small (Sm)`, `Medium (Me)`, and `Big (Bi)`. The expressions of the basic trichotomy may be modified using linguistic hedges either with *narrowing* or *widening* effect. The hedges with narrowing effect, ordered according to the narrowing effect, are `Very` (Ve), `Significantly` (Si) and `Extremely` (Ex). The hedges with widening effect, ordered according to the narrowing effect, are `More or Less` (ML), `Roughly` (Ro) and `Quite Roughly` (QR).

Such linguistic expressions have their theoretical model of semantics based on intension, context, and extension, which is in detail described in the referred literature [21]. For the purpose of this contribution, it is sufficient to mention that extensions, that model the meaning in a given context $[v_L, v_R]$, are fuzzy sets that are depicted in Figure 1. One may see the influence of the modifiers on the shape of the extensions.

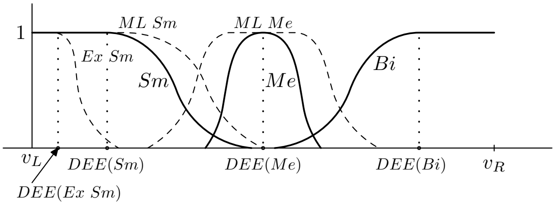


Fig. 1. Shapes of extensions (fuzzy sets) of evaluative linguistic expressions

If a fuzzy rule base is viewed as a linguistic description, and thus uses the above recalled evaluative linguistic expressions with their model of semantics, one can neither model them as a conjunction of implicative rules nor as a disjunction of conjunctions (Mamdani-Assilian model). The used expressions, mainly the full inclusion of their models (fuzzy sets), require a specific inference method – *Perception-based Logical Deduction* (PbLD) [20]. This method models each fuzzy rule

$$\mathcal{R}_i := \mathsf{IF}\ \mathtt{X}\ \text{is}\ \mathcal{A}_i\ \mathsf{THEN}\ \mathtt{Y}\ \text{is}\ \mathcal{B}_i,$$

by a fuzzy relation R_i on $X \times Y$ given as follows:

$$R_i(x, y) = A_i(x) \rightarrow_{\text{Ł}} B_i(y), \quad x \in X, y \in Y$$

where $\rightarrow_{\text{Ł}}$ is the Łukasiewicz implication given by $a \rightarrow_{\text{Ł}} b = 1 \wedge (1 - a + b)$. For the sake of clarity, let us note that X, Y denote the so called linguistic variable that take values from a set of linguistic expressions, these linguistic expressions are modelled by fuzzy sets (extensions) on given universes (contexts) X, Y, and finally, $x \in X$ and $y \in Y$.

However, unlike in the case of implicative rules, the rules are not aggregated conjunctively. The PbLD uses a specific algorithm (perception) that chooses only some rules to be used in the inference, particularly, the most specific ones among the most fired rules. Only outputs obtained based on these fuzzy rules are finally aggregated by the intersection. For details, we refer to [7,25].

Finally, the inferred output is defuzzified by the *Defuzzification of Evaluative Expressions* (DEE) that has been designed specifically for the outputs of the PbLD inference mechanism. In principle, DEE is a combination of *First-Of-Maxima* (FOM), *Mean-Of-Maxima* (MOM) and *Last-Of-Maxima* (LOM) that are applied based on the classification of the inferred output fuzzy set. If the inferred fuzzy set is of the type `Small`, the LOM is applied; if the inferred output is of the type `Medium`, the MOM is applied; and finally, if the inferred output is of the type `Big`, the FOM is applied, see Figure 1. In the case of the FRBE method, the defuzzification DEE is applied after the inference, so that the deduced weights are already crisp numbers.

2.3 Fuzzy Rule Base Identification

The last missing point is the identification of the linguistic descriptions. This may be done by distinct approaches. One could expect a deep applicable expert knowledge, however, neither our experience nor the experience of others confirms this expectations. Let us once more refer to the observation of Armstrong, Collopy, and Adya in [2], already recalled in Section 1.

Because of the missing reliable expert knowledge, we focus on data-driven approaches that may bring us the interpretable knowledge hidden in the data.

However, before we apply any data-mining technique, we have to clarify how we interpret the weights in the data. Naturally, the individual method weights should be proportionally higher if a given method is supposed to provide lower forecasting error and vice-versa. Thus, it is natural to put

$$w_j = 1 - acc_j, \quad j = 1, \dots, M$$

where acc_j denotes an appropriate normalized forecasting error of the j-th method. Now, any appropriate data-mining technique may be applied in order to determine the dependence between features and the weight of each method.

3 Fuzzy GUHA – Linguistic Associations Mining

In this paper, we employ the so called linguistic associations mining for the fuzzy rule base identification. This approach, mostly known as mining association rules [1] and firstly introduced as GUHA method [8,9], finds distinct statistically approved associations between attributes of given objects. Particularly, the GUHA method deals with Table 1 where $o_1, \dots, o_n$ denote objects, $X_1, \dots, X_m$ denote independent boolean attributes, Z denotes the dependent (explained) boolean attribute, and finally, symbols a_{ij} (or a_i) $\in \{0, 1\}$ denote whether an object o_i carries an attribute X_j (or Z) or not.

Table 1. Standard GUHA Table

	X_1	$\dots$	X_m	Z
o_1	a_{11}	$\dots$	a_{1m}	a_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
o_n	a_{n1}	$\dots$	a_{nm}	a_n

The original GUHA allowed only boolean attributes to be involved [10]. Since most of the features of objects are measured on the real interval, standard approach assumed to binarize the attributes by a partition of the interval into subintervals. The goal of the method is to search for associations of the form

$$\mathtt{A}(X_1, \ldots, X_p) \simeq \mathtt{B}(Z)$$

where $\mathtt{A}$, $\mathtt{B}$ are predicates containing only the connective $\mathtt{AND}$ and $X_1, \ldots, X_p$ for $p \leq m$ are all variables occurring in $\mathtt{A}$. The $\mathtt{A}$, $\mathtt{B}$ are called the *antecedent* and *consequent*, respectively.

The relationship between the antecedent and consequent is described by the so called *quantifier* $\simeq$. There are many quantifiers that characterize validity of the association in data [9]. For our task, we use the so called *binary multitudinal quantifier* $\simeq := \sqsubset_r^\gamma$. Let a denotes the number of positive occurrences of $\mathtt{A}$ as well as $\mathtt{B}$ in the data; let b be the number of positive occurrences of $\mathtt{A}$ and of negated $\mathtt{B}$, i.e. of 'not $\mathtt{B}$'. Then the above mentioned quantifier is taken as true if

$$\frac{a}{a+b} > \gamma \qquad \text{and} \qquad \frac{a}{n} > r,$$

where $\gamma \in [0, 1]$ is a *degree of confidence* and $r \in [0, 1]$ is a *degree of support.*

In many situations, including ours, the fuzzy variant of the GUHA method [14,22] seems to be more appropriate. We adopt the variant first used in [26] where the attributes are not boolean but vague, particularly expressed by means of evaluative linguistic expressions. With three basic expressions `Small, Medium, Big` and seven different linguistic hedges (including the empty one), we define 18 fuzzy sets for every quantitative variable (hedges with narrowing effect and expression `Medium` are omitted). The values a_{ij} (or a_i) are elements of the interval $[0, 1]$ that express membership degrees to these fuzzy sets.

The binary multitudinal quantifier is constructed analogously to the one in crisp GUHA. The difference is that the numbers a, b are not summations of 1s and 0s, but summations of membership degrees of objects into fuzzy sets representing the antecedent $\mathtt{A}$ and consequent $\mathtt{B}$, or its complement, respectively. Naturally, the fact, that antecedent $\mathtt{A}$ as well as consequent $\mathtt{B}$ hold simultaneously, leads to the natural use of a *t-norm*. In our case, we use the Gödel t-norm, i.e., the minimum. For example, if an object o_i belongs to a given antecedent in a degree 0.7 and to a given consequent in a degree 0.6, the value that enters the summation equals to $\min\{0.7, 0.6\} = 0.6$. Summation of such values over all the objects equals to the value a, the other value b is determined analogously. The rest of the ideas of the method remain the same.

By using fuzzy sets, we generally get more precise results, and, more importantly, we avoid undesirable threshold effects [24]. The further advantage is that the method searches for implicative associations that may be directly interpreted as fuzzy rules for the PbLD inference system. In our case, for each individual forecasting method, we have transformed the training data set of time series with their normalized features into a table similar to Table 2.

The rest of this section deals with ARIMA method. Of course, the same process has been applied for all the other forecasting methods in our ensemble.

Table 2. Transformed Training Data Set for the ARIMA Forecasting Method

	Φ_1^{ExSm}	$\ldots$	Φ_q^{ExBi}	$W_{\text{AR}}^{\text{ExSm}}$	$\ldots$	$W_{\text{AR}}^{\text{ExBi}}$
TS_1	0.9	$\ldots$	0.7	0	$\ldots$	0.9
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
TS_n	0.1	$\ldots$	0.2	0.8	$\ldots$	0

Objects $\text{TS}_1, \ldots, \text{TS}_n$ in Table 2 are the time series from the training set; $\Phi_1, \ldots, \Phi_q$ are normalized features of given time series. Note that there are significantly more columns in this part of Table 2 because each evaluative linguistic expression leads to a single column for a single feature Φ_i, i.e. for the expression `ExSm`, there are q columns: $\Phi_1^{\text{ExSm}}, \ldots, \Phi_q^{\text{ExSm}}$, where q denotes the number of features. Once more, let us recall that we construct 18 linguistic expressions.

Symbol W_{AR} stands for the weight (inverted accuracy) of the ARIMA method, and again, there are as many columns in this part of the Table 2 as there exist evaluative linguistic expressions, i.e. 18 in the chosen setting. The fuzzy GUHA then combinatorically generates hypotheses that are immediately statistically either declined or confirmed as linguistic associations based on the chosen quantifier parameters. For our purposes, based on a set of experiments, we set up the thresholds for $\gamma = 0.65$ and $r = 0.05$.

Note that the above described application of the fuzzy GUHA method generates linguistic description determining the weight of a single method – in our example of the ARIMA method. Thus, the method, including the transformation of training data set into a table similar to Table 2, has to be applied as many times as is the number of methods (and consequently of the linguistic descriptions). In our case, this led to the fourfold use of the method as we deal with four individual forecasting methods.

4 Implementation

To develop and validate the model, we have used 2829 time series from the M3 data set repository that contains 3003 time series from the M3-Competition [17]. We have omitted time series with other than yearly, quarterly, and monthly frequencies. Note, that the M3 set of time series serves as a generally accepted benchmark database provided by the authority of the International Institute of Forecasters. This selected data set was divided into two distinct sets simply by putting time series with even or odd IDs into the *training set* and the *testing set*, respectively.

The training set was used for an identification of our model, that is, for generation of our fuzzy rule base. The testing set was used for testing whether the determined knowledge encoded in the fuzzy rules works generally also for time series “not seen” by the rule base generating GUHA algorithm.

All forecasts were computed with the R software, version 3.1.0 beta, and package `forecast` version 5.3 [13]. We have chosen the often used forecasting methods: *seasonal Autoregressive Integrated Moving Average* (R-ARIMA), *Exponential Smoothing* (R-ES), *Random Walk process* (R-RW) and *Theta* (R-Theta). For details about these methods, we only refer to the relevant literature [3,5,11,18].

These methods were executed with fully automatic parameter selection and optimization which made it possible to concentrate the investigation purely on the combination technique. Moreover, their arithmetic mean (R-AM), i.e., the equal weights ensemble method, was also determined and used as a benchmark.

There are many accuracy measures that are used to analyze the performance of the various forecasting methods. However, very popular measures such as *Mean Absolute Error* or *(Root) Mean Squared Error* are inappropriate for comparison across more time series because they are scale-dependent. We use *Symmetric Mean Absolute Percentage Error* (SMAPE) that is scale-independent and thus, appropriate in order to compare methods across different time series [12].

Let a given time series $y_1, y_2, \ldots, y_T$ be of the frequency F, i.e. $F = 1, 4, 12$, for yearly, quarterly, and monthly time series, respectively. Based on experiments and previous publications [15], the following features were considered in introductory studies [23,26] as well as in this paper.

The normalized *frequency* is given by the reciprocal value of F, i.e., it is given as 1/12, 1/4, and 1 in case of the monthly, quarterly, and yearly time series, respectively. The normalized *length of the time series* is given by $\min(T/100, 1)$ where T denotes the number of known time lags. Further, the *skewness*, the *kurtosis* and the *coefficient of variation* as standard statistics are also normalized and taken into account. Finally, the *strength of trend*, *strength of seasonality* and the *stationarity* are also considered. These features are obtained as $(1 - p)$ values, where p is a p-value of an appropriate statistical test, e.g., the Augmented Dickey–Fuller test in the case of stationarity.

5 Results

As mentioned above, the associations generated by GUHA method are implicative. Thus, they may be directly interpreted as fuzzy rules. Due to the large amount of such generated rules, a redundancy removal [7,25] and size reduction algorithms were applied on these rules, which significantly reduced the numbers of rules.

In order to judge its performance, the fuzzy rule-based ensemble was applied on 1415 time series from the testing set, i.e. on all monthly, quarterly and yearly times series with odd IDs in the M3 competition. Table 3 shows that arithmetic mean and standard deviation of SMAPE forecasting errors over all testing time series is better for fuzzy rule-based ensemble than any individual forecasting method from the R package used in the ensemble. Moreover, the equal-weights, i.e. arithmetic mean (R-AM), and the three best methods from the M3 competition according to the average precision on the testing set (M3-THETA, M3-ForecastPro, M3-ForcX) have been outperformed as well.

To indicate superiority of our method, a statistical test of significance has been performed. Namely, we have performed Wilcoxon signed rank test with continuity correction for the null hypothesis that the median of the random variable

Table 3. Average and Standard Deviation of the SMAPE Forecasting Errors

Method	Error Average	Error Std.Dev.
FRBE	**13.29**	**14.05**
M3-THETA	13.56	15.42
R-AM	13.66	14.22
M3-ForecastPro	13.67	15.50
M3-ForcX	13.76	15.26
R-ES	13.95	15.23
R-ARIMA	14.58	16.77
R-THETA	14.73	15.33
R-RW	16.53	17.20

($\mathrm{SMAPE}_{\text{R-method}} - \mathrm{SMAPE}_{\mathrm{FRBE}}$) equals to zero, with the non-zero equality alternative hypothesis. The null hypothesis was rejected for all methods from R including the R-AM in the standard significance level $\alpha = 0.05$. Particularly, the obtained p-value adjusted for multiple comparisons was less than 1.80×10^{-3} for R-ES, less than 5.41×10^{-7} for R-ARIMA, less than 1.38×10^{-24} for R-AM, etc. Similar hypotheses could be rejected for M3-ForcX only on the significance level $\alpha = 0.10$ as the p-value equaled to 8.22×10^{-2} but could not be rejected for the other two M3 methods.

Let us stress that the best performance has been reached also in the robustness (standard deviation of the SMAPE forecasting errors, see Table 3), which is perhaps even more important w.r.t. the goals of ensemble methods. To compare variances of $\mathrm{SMAPE}_{\mathrm{AM}}$ and $\mathrm{SMAPE}_{\mathrm{FRBE}}$, the F-test was performed. As a result, null hypothesis of ratio of variances being equal to 1 was rejected for all methods excepting for the R-AM where the adjusted p-value was equal to 0.66.

6 Critical Discussion and Future Directions

The obtained results showed an improvement in the accuracy as well as in the standard deviation of the accuracy that confirms the improvement in the sense of "robustness". Let us now open a short discussion related to the results and the approach. Undoubtedly, the results confirm some sort of improvement. One could surely express objections to the too slight improvement and also to the too difficult and technologically demanding approach.

Related to the first objection, we have to stress that we have tested the improvement in accuracy not only compared to the arithmetic mean but also compared to all the individual methods (with p-value adjustment for multiple comparisons). The suggested FRBE method was found significantly better in median error than all the used individual methods in R including the equal weights method, which confirms the significantly positive influence of the ensemble. Moreover, the ensemble, composed only from the standard methods included in R, outperformed all three top M3 methods although not significantly. The null hypothesis of the variance F-test was rejected in all cases with the only exception of R-AM. In other words, all the used methods, either those participating on the ensemble or those best methods according to M3 competition, were significantly outperformed either in precision, or in robustness, or in both criteria.

As a future direction, we plan to employ a stochastic optimization task implemented on a high performance computer in order to find the optimal setting of all "bricks" building the FRBE. This does not relate only to the individual methods, but also to the features itself, and their normalization. For example, the used $(1-p)$-values (*strength of trend, strength of seasonality, stationarity*) lie in the $[0,1]$-interval and thus, are not further normalized anymore. However, $(1-p)$-values around 0.7 or 0.8 are extremely low from the statistical point of view, as p-value around 0.2 usually does not allow to reject null hypothesis. But within the standard context $[0,1]$, the values around 0.8 are found rather big. Narrowing the interval of p-values and consequently the derived features given by $(1-p)$-values seems to be necessary. Nevertheless, the particular realization of the narrowed normalization is again an open question that may be solved within the more general optimization task performed by the stochastic optimization implemented on a supercomputer.

Regarding the second objections, let us stress that the difficulty appears only in the construction phase. In the final phase, that is planned to be reached, we expect to have a rather simple (from a user point of view) tool that will automatically determine a given time series features, use the pre-determined fuzzy rules to set-up weights of individual methods, perform individual method forecasts, combine them according to the determined weights, and finally, provide a user with a single accurate yet robust forecast.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. on Very Large Databases, pp. 487–499. AAAI Press, Chile (1994)
2. Armstrong, J.S., Adya, M., Collopy, F.: Rule-based forecasting using judgment in time series extrapolation. In: Armstrong, J.S. (ed.) Principles of Forecasting: A Handbook for Reasearchers and Practitioners. Kluwer Academic Publishers, Boston (2001)
3. Assimakopoulos, V., Nikolopoulos, K.: The theta model: a decomposition approach to forecasting. International Journal of Forecasting 16(4), 521–530 (2000)
4. Bates, J.M., Granger, C.W.J.: Combination of forecasts. Operational Research Quarterly 20, 451–468 (1969)
5. Box, G., Jenkins, G.: Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1976)
6. Collopy, F., Armstrong, J.S.: Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. Management Science 38, 1394–1414 (1992)
7. Dvořák, A., Štěpnička, M., Vavříčková, L.: Redundancies in systems of fuzzy/linguistic if-then rules. In: Proc. 7th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011) and LFA-2011. Advances in Intelligent Systems Research, pp. 1022–1029. Atlantic Press, Paris (2011)

8. Hájek, P.: The question of a general concept of the GUHA method. Kybernetika 4, 505–515 (1968)
9. Hájek, P., Havránek, T.: Mechanizing hypothesis formation: Mathematical foundations for a general theory. Springer, Heidelberg (1978)
10. Hájek, P., Holeňa, M., Rauch, J.: The GUHA method and its meaning for data mining. Journal of Computer and Systems Sciences 76, 34–48 (2010)
11. Hamilton, J.D.: Time Series Analysis. Princeton University Press, New Jersey (1994)
12. Hyndman, R., Koehler, A.: Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688 (2006)
13. Hyndman, R.J., Athanasopoulos, G., Razbash, S., Schmidt, D., Zhou, Z., Khan, Y., Bergmeir, C.: forecast: Forecasting functions for time series and linear models (2014), `http://CRAN.R-project.org/package=forecast` (r package version 5.3)
14. Kupka, J., Tomanová, I.: Some extensions of mining of linguistic associations. Neural Network World 20, 27–44 (2010)
15. Lemke, C., Gabrys, B.: Meta-learning for time series forecasting in the nn gc1 competition. In: Proc. 16th IEEE Int. Conf. on Fuzzy Systems, Barcelona, pp. 2258–2262 (2010)
16. Makridakis, S., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R.: The accuracy of extrapolation (time-series) methods - results of a forecasting competition. Journal of Forecasting 1, 111–153 (1982)
17. Makridakis, S., Hibon, M.: The m3–competition: results, conclusions and implications. International Journal of Forecasting 16, 451–476 (2000)
18. Makridakis, S., Wheelwright, S., Hyndman, R.: Forecasting: methods and applications. John Wiley & Sons, USA (2008)
19. Newbold, P., Granger, C.W.J.: Experience with forecasting univariate time series and combination of forecasts. Journal of the Royal Statistical Society Series a-Statistics in Society 137, 131–165 (1974)
20. Novák, V.: Perception-based logical deduction. In: Reusch, B. (ed.) Computational Intelligence, Theory and Applications. ASC, pp. 237–250. Springer, Heidelberg (2005)
21. Novák, V.: A comprehensive theory of trichotomous evaluative linguistic expressions. Fuzzy Sets and Systems 159(22), 2939–2969 (2008)
22. Novák, V., Perfilieva, I., Dvořák, A., Chen, Q., Wei, Q., Yan, P.: Mining pure linguistic associations from numerical data. International Journal of Approximate Reasoning 48, 4–22 (2008)
23. Sikora, D., Štěpnička, M., Vavříčková, L.: Fuzzy rule-based ensemble forecasting: Introductory study. In: Kruse, R., Berthold, M., Moewes, C., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Synergies of Soft Computing and Statistics. AISC, vol. 190, pp. 379–387. Springer, Heidelberg (2013)
24. Sudkamp, T.: Examples, counterexamples, and measuring fuzzy associations. Fuzzy Sets Systems 149(1), 57–71 (2005)
25. Štěpničková, L., Štěpnička, M., Dvořák, A.: New results on redundancies of fuzzy/linguistic if-then rules. In: Proc. 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2013), pp. 400–407. Atlantic Press, Milano (2013)
26. Štěpničková, L., Štěpnička, M., Sikora, D.: Fuzzy rule-based ensemble with use linguistic associations mining for time series prediction. In: Proc. 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2013), pp. 408–415. Atlantic Press, Milano (2013)

Multistage Fuzzy Control of a Stochastic System Using a Bacterial Genetic Algorithm

Janusz Kacprzyk

Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01–447 Warsaw, Poland
kacprzyk@ibspan.waw.pl

Abstract. We consider multistage control problem under fuzzy constraints on controls applied and fuzzy goals on states attained, with a stochastic system under control (a Markov chain). We seek an optimal sequence of controls which maximizes the probability of attaining the fuzzy goal subject to the fuzzy constraints, over a finite, fixed and specified planning horizon. We present an extension of Kacprzyk's [10,12] approach, based on a traditional genetic algorithm, by employing a bacterial evolutionary algorithm in the setting of Nawa and Furuhashi [18]. We show that it yields an improved efficiency, and potentials for future extensions.

Keywords: fuzzy control, multistage fuzzy control, fuzzy dynamic programming, stochastic system under control, genetic algorithm, pseudo-bacterial genetic algorithm, bacterial evolutionary algorithm.

1 Introduction

We consider multistage fuzzy optimal control under fuzzy constraints on inputs (controls) and fuzzy goals on outputs (states attained) in the setting of Bellman and Zadeh [1], comprehensibly extended by Kacprzyk, notably in his books [6,9].

In the basic case of a deterministic system under control, given as a state transition equation $x_{t+1} = f(x_t, u_t)$, $t = 0, 1, \ldots, N-1, \leq \infty$, where $x_t, x_{t+1} \in X = \{s_1, \ldots, s_n\}$ are the states (outputs) at control stages t and $t+1$, respectively, and $u_t \in U = \{c_1, \ldots, c_m\}$ is the control at control stage t, at each control stage t, $t = 0, 1, \ldots, N-1$, $u_t \in U$ is subjected to a *fuzzy constraint*, $\mu_{C^t}(u_t)$, and on $x_N \in X$ a *fuzzy goal*, $\mu_{G^N}(x_N)$ is imposed; N is the termination time which is fixed and specified in advance. The initial state $x_0 \in X$ is known in advance.

The fuzzy decision (performance function) is

$$\mu_D(u_0, \ldots, u_{N-1} \mid x_0) = \mu_{C^0}(u_0) \wedge \cdots \wedge \mu_{C^{N-1}}(u_{N-1}) \wedge \mu_{G^N}(x_N) \qquad (1)$$

and the problem is as to find an optimal sequence of controls $u_0^*, \ldots, u_{N-1}^*$, $u_t^* \in U$, $t = 0, 1, \ldots, N-1$, such that

$$\mu_D(u_0^*, \ldots, u_{N-1}^* \mid x_0) = \max_{u_0, \ldots, u_{N-1} \in U} \mu_D(u_0, \ldots, u_{N-1} \mid x_0) \qquad (2)$$

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_32

We assume here a stochastic system under control with state transitions given by the conditional probability $p(x_{t+1} \mid x_t, u_t)$; $t = 0, 1, \dots, N-1$, and show the problem formulation in Section 2, and its solution by dynamic programming. In Section 3 we outline the solution by a traditional genetic (GA) algorithm due to Kacprzyk [8], [10], [12]. In Section 4 we propose the solution by a bacterial evolutionary algorithm (BEA) due to Nawa and Furuhashi [18]. We show its good efficiency vs. the ordinary GA. We provide conclusions and point out potentials, notably using a novel concept of a memetic bacterial algorithm (MBA) due to by Kóczy and his collaborators (cf. [2], [3], [4]).

2 Multistage Control of a Stochastic System in a Fuzzy Environment

We deal with the control problem (2) with a stochastic system under control, i.e. with a *joint occurrence* of fuzziness and randomness (cf. [7] for a general review).

The stochastic system under control is a *Markov chain* with dynamics (state transitions) governed by a conditional probability function $p(x_{t+1} \mid x_t, u_t)$, $t = 0, 1, \dots, N$, $N \leq \infty$, which specifies the probability of attaining $x_{t+1} \in X = \{s_1, \dots, s_n\}$ from $x_t \in X$, under $u_t \in U = \{c_1, \dots, c_m\}$.

At each $t = 0, 1, \dots, N-1$, $u_t \in U$ is subjected to a fuzzy constraint $\mu_{C^t}(u_t)$, and on $x_N \in X$ a fuzzy goal $\mu_{G^N}(x_N)$ is imposed. The value of $\mu_D(u_0, \dots, n_{N-1} \mid x_0)$ is a random variable, and we employ the expected value in the problem formulation.

Basically the two different problem formulations are used:

1. due to Bellman and Zadeh's [1], that is: we seek $u_0^*, \dots, u_{N-1}^*$ such that $u_0^*, \dots, u_{N-1}^*$ that

$$\begin{aligned}\mu_D(u_0^*, \dots, u_{N-1}^* \mid x_0) = \max_{u_0, \dots, u_{N-1}} [\mu_{C^0}(u_0) \wedge \dots \\ \dots \wedge \mu_{C^{N-1}}(u_{N-1}) \wedge E\mu_{G^N}(x_N)] \qquad (3)\end{aligned}$$

2. due to Kacprzyk and Staniewski [14] (cf. Kacprzyk [9]), that is, we seek $u_0^*, \dots, u_{N-1}^*$ such that

$$\begin{aligned}\mu_D(u_0^*, \dots, u_{N-1}^* \mid x_0) = \max_{u_0, \dots, u_{N-1}} E\mu_D(u_0, \dots, u_{N-1} \mid x_0) = \\ = \max_{u_0, \dots, u_{N-1}} E[\mu_{C^0}(u_0) \wedge \dots \wedge \mu_{C^{N-1}}(u_{N-1}) \wedge \mu_{G^N}(x_N)] \qquad (4)\end{aligned}$$

We will employ the classic Bellman and Zadeh's [1] formulation (3) which is more commonly used, and better suits our purpose.

First, G^N is regarded as a fuzzy event in X, and the conditional probability of G^N given x_{N-1} and u_{N-1} is

$$E\mu_{G^N}(x_N) = E\mu_{G^N}(x_N \mid x_{N-1}, u_{N-1}) = \sum_{x_N \in X} p(x_N \mid x_{N-1}, u_{N-1}) \cdot \mu_{G^N}(x_N) \qquad (5)$$

Clearly, the structure of problem (3) makes the use of dynamic programming possible (cf. Kacprzyk [9]), and the dynamic programming recurrence equations are:

$$\begin{cases} \mu_{G^{N-1}}(x_{N-1}) = \max_{u_{N-1}}[\mu_{C^{N-i}}(u_{N-i}) \wedge E\mu_{G^{N-i+1}}(x_{N-i+1})] \\ E\mu_{G^{N-1+1}}(x_{N-i+1}) = \sum_{x_{N-i} \in X} p(x_{N-i+1} \mid x_{N-i}, u_{N-i}) \times \mu_{G^{N-i+1}}(x_{N-i+1}) \\ i = 1, \ldots, N \end{cases} \tag{6}$$

The successive maximizing values of u_{N-i}, u^*_{N-i}, $i = 1, 2, \ldots, N$, give the optimal control policies $a^*_{N-i} : X \longrightarrow U$ such that $u^*_{N-i} = a^*_{N-i}(x_{N-i})$, $i = 1, \ldots, N$.

Though dynamic programming finds an optimal solution to (3), it suffers the from known infamous curse of dimensionality – cf. Kacprzyk [9]. Therefore, Kacprzyk [8,12] proposed the use of a genetic algorithm which proved to be conceptually simple and numerically efficient,and will now be outlines to provide a point of departure for this paper.

3 Using a Genetic Algorithm for the Multistage Fuzzy Control of a Stochastic System

This essence of Kacprzyk's [8,10,12] approach, for the stochastic system considered, is as follows. By an *individual* we mean a particular solution, values of controls at the consecutive control stages, $u_0, \ldots, u_{N-1}$. It is *evaluated* by the fuzzy decision [maximized in 3], the *fitness function.* The *population* is a set of potential solutions, here of a fixed size. We initially assume some initial population which is randomly generated. Then, some members of the population, the parents, undergo *reproduction* through *crossover* and *mutation* to produce off-springs (children), i.e. some new solutions. Then, the best ones (the fittest) "survive", i.e. are used while repeating this process. Finally, at the end of such a process one may expect to find a very good (if not optimal) solution.

More formally:

- the problem is represented by strings of controls $u_0, \ldots, u_{N-1}$, and we use real coding;
- the fitness (evaluation) function is the fuzzy decision, i.e.

$$\mu_D(u_0, \ldots, u_{N-1} \mid X_0) = \mu_{C^0}(u_0) \wedge \mu_{C^{N-1}}(u_{N-1}) \wedge E\mu_{G^N}(x_N) \tag{7}$$

- a standard random selection, crossover and mutation, and a standard termination condition, mainly a predefined number of iterations, or iteration-to-iteration improvement lower than a threshold, are used.

For the problem class considered, we assume the following granulation: the state and control spaces are real intervals, i.e. - respectively - $X = [s_1, s_n], s_1, s_2, \ldots, s_n \in R$, $s_1 < s_2 < \ldots < s_n$; $U = [c_1, c_m], c_1, c_2, \ldots, c_m \in R$, $c_1 < c_2 < \ldots < c_m$; n and

m are properly chosen, and the partitioning $s_1, \ldots, s_n$ and $c_1, \ldots, c_m$ need not be evenly spaced over $[s_1, \ldots, s_n]$ and $[c_1, \ldots, c_m]$,respectively.

The genetic operations in our case are standard: if we have two individuals (solutions), i.e. strings of controls, $u_0, u_1, \ldots, u_{N-1}$, i.e.

$$\begin{cases} S_1 = (u_0^1, u_1^1, \ldots, u_{k-1}^1, u_k^1, u_{k+1}^1, \ldots, u_{m-1}^1, u_m^1, u_{m+1}^1, \ldots u_{N-1}^1) \\ S_2 = (u_0^2, u_1^2, \ldots, u_{k-1}^2, u_k^2, u_{k+1}^2, \ldots, u_{m-1}^2, u_m^2, u_{m+1}^2, \ldots u_{N-1}^2) \end{cases} \tag{8}$$

then we randomly generate two points, k and m, $1 < k \leq m < N-1$, and apply the classic two-point crossover to generate the two off-springs

$$\begin{cases} S_1^a = (u_0^1, u_1^1, \ldots, u_{k-1}^1, u_k^2, u_{k+1}^2, \ldots, u_{m-1}^2, u_m^1, u_{m+1}^1, \ldots u_{N-1}^1) \\ S_2^b = (u_0^2, u_1^2, \ldots, u_{k-1}^2, u_k^1, u_{k+1}^1, \ldots, u_{m-1}^1, u_m^2, u_{m+1}^2, \ldots u_{N-1}^2) \end{cases} \tag{9}$$

As for the mutation, we apply the dynamic non-uniform mutation (cf. Herrera, Lozano and Verdegay [5]), that is, if we have a solution $S = (u_0, u_1, \ldots, u_k, \ldots, u_{N-1}^1)$, and we randomly select the gene u_k (control at stage k) to be mutated, and randomly generate a number $v \in \{-1, 0, 1\}$, then the mutated gene, $\overline{u}_k$, is

$$\overline{u}_k = u_k + v \times \delta \tag{10}$$

where δ is some small value from $U = [c_1, c_m]$. Therefore, (10) slightly changes the control at $t = k$ in an individual, to the nearest $c_k \in \{c_1, \ldots, c_m\}$. Moreover, optionally we can dynamically change δ in (10) so that its value in early iterations may be higher than at later ones, e.g. in the spirit of (Michalewicz and Janikow[16]).

The genetic algorithm works is now basically:

```
begin
  $t := 0$
  set the initial population $P(t)$ consisting of randomly generated strings
      of controls (i.e. of randomly generated real numbers from $[0, 1]$);
  for each $u_0, \ldots, u_{N-1}$, in each string in $P(t)$, find the resulting $x_{t+1}$ by using
      the state transition equation $x_{t+1} = f(x_t, u_t)$, and use (1) to evaluate each
      string in $P(t)$;
  while $t <$ maximum number of iterations do
  begin
      $t =: t + 1$
      assign the probabilities to each string in $P(t-1)$ which are proportional
      to the value of (1) for each string;
      randomly (using those probabilities) generate the new population $P(t)$;
      perform crossover and mutation on the strings in $P(t)$;
      calculate the value of (1) for each string in $P(t)$.
  end
end
```

For more details on this genetic algorithms, and some “trickery” applied, cf. Kacprzyk [12].

4 Using a Bacterial Evolutionary Algorithm (BEA) for the Multistage Fuzzy Control of a Stochastic System

Though the use of a genetic algorithm for the solution of the problem considered outlined in the previous section has proved to be conceptually simple, and quite effective and efficient, in this paper we extend that approach by employing a novel approach, a bacterial evolutionary algorithm (BEA), originally proposed for fuzzy rule extraction and optimization of a much more general applicability, notably for our the problem.

Due to lack of space we will only outline the idea of BEA. We should however start with its predecessor, the *pseudo-bacterial genetic algorithm* (PBGA) proposed by Nawa, Hashiyama, Furuhashi and Uchikawa [17]. Its idea boils down to a new genetic operation called a *bacterial mutation* mimicking bacterial evolution which intends to improve parts of chromosomes contained in each bacteria using a mechanism of transferring genes to other bacteria. The first step is to determine how the problem can be encoded in a particular bacteria (chromosome). In our case, these are the particular values of controls over the consecutive stages. i.e. given by (8).

Then, the pseudo-code of a PBGA can roughly written as:

```
begin
  $t =: 0$
  set the initial population $N_{Ind}$ consisting of randomly generated
    strings of controls (i.e. of randomly generated real numbers from $[c_1, c_m]$,
    taking the nearest $c_{\{.\}}$);
  for each $u_0, \ldots, u_{N-1}$,
    evaluate each string (8) in $N_{Ind}$ using (1)
    create clones (clones) of the solutions selected;
  while all clones are mutated and tested exactly once do
    apply the bacterial mutation to each solution (string of controls) selected
      consecutively;
    choose the same copies of the selected solutions ("clones");
    choose the same part or parts randomly from the clones and mutate it
      (except for one clone that is unchanged);
    select the best clone and transfer its mutated part or parts
      to the other clones;
  end;
  leave the best clones only and remove other ones;
  while $t <$ maximum number of iterations do
  begin – GA STEP
    $t := t + 1$
    assign the probabilities to each string remaining in $N_{Ind}$ which are
      proportional to the value of (1) for each string;
    randomly (using those probabilities) generate the new population $P(t)$;
```

```
        perform crossover and mutation on the strings in $P(t)$;
        calculate the value of (1) for each string in $P(t)$.
    end
end
```

Though the above PBGA algorithm works quite well, also for the solution of our problem, many authors have reported good results with a further extension which results in a *bacterial evolutionary algorithm* (BEA) proposed by Nawa and Furuhashi [18] in which, basically, a new operation is added to the PBGA called a *gene transfer operation* which sets relationships between solutions in the population. Briefly, in in the above pseudo-code of the PBGA, in **GA STEP** we use instead of the usual selection, crossover and mutation the following gene transfer operations:

- Sort the population of solutions according to the fitness values (1) and divide it in two halves: of better (superior half) and worse (inferior half) solutions;
- Choose one solution (the "source chromosome") from the superior half and another one (the "destination chromosome") from the inferior half;
- Transfer a part (selected randomly) from the source chromosome to the destination chromosome;
- Repeat the steps above N_{Ind} times.

The algorithm is conceptually simple and easily implementable if we already have, as we do, an implemented GA a PBGA algorithms for our problem.

5 Application of the Bacterial Evolutionary Algorithm for the Multistage Fuzzy Control of a Stochastic System

We will illustrate now this algorithm by a simple example, showing first the results obtained by employing the traditional GA (cf. Kacprzyk [12]), and comparing them with those obtained for the new BEA.

Suppose that the state space is $X = \{s_1, \ldots, s_{10}\}$, the control space is $U = \{c_1, \ldots, c_8\}$, the planning horizon is $N = 10$, and the initial state is $x_0 = s_1$. The state transitions are governed by the following conditional probability [notice that the particular values of u_t and x_t correspond to the rows while those of x_{t+1} to the columns]:

$$p(x_{t+1} \mid x_t, u_t) =$$

		s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
$u_t = c_1$	s_1	0.0	0.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	s_2	0.0	0.0	0.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0
	...	...	...	...	...	...	...	...	...	...	...
	s_{10}	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

		s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
$u_t = c_2$	s_1	0.0	0.1	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	s_2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.8	0.1	0.0
	...	...	...	...	...	...	...	...	...	...	...
	s_{10}	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

		s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
$u_t = c_8$	$x_t = s_1$	0.0	0.0	0.0	0.0	0.5	0.4	0.1	0.0	0.0	0.0
	s_2	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.2	0.0	0.0
	...	...	...	...	...	...	...	...	...	...	...
	s_{10}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.6

The fuzzy constraints and goal are:

$$
\begin{aligned}
C^0 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 0.6/c_4 + 0.3/c_5 + 0.1/c_6 + 0.0/c_7 + 0.0/c_8\\
C^1 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 0.6/c_4 + 0.3/c_5 + 0.1/c_6 + 0.0/c_7 + 0.0/c_8\\
C^2 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 0.7/c_4 + 0.5/c_5 + 0.3/c_6 + 0.2/c_7 + 0.1/c_8\\
C^3 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 0.9/c_4 + 0.7/c_5 + 0.5/c_6 + 0.4/c_7 + 0.2/c_8\\
C^4 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 1.0/c_4 + 0.9/c_5 + 0.6/c_6 + 0.5/c_7 + 0.4/c_8\\
C^5 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 1.0/c_4 + 1.0/c_5 + 0.7/c_6 + 0.6/c_7 + 0.5/c_8\\
C^6 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 1.0/c_4 + 1.0/c_5 + 0.8/c_6 + 0.7/c_7 + 0.6/c_8\\
C^7 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 1.0/c_4 + 1.0/c_5 + 0.9/c_6 + 0.8/c_7 + 0.7/c_8\\
C^8 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 1.0/c_4 + 1.0/c_5 + 1.0/c_6 + 0.9/c_7 + 0.8/c_8\\
C^9 &= 1.0/c_1 + 1.0/c_2 + 1.0/c_3 + 1.0/c_4 + 1.0/c_5 + 1.0/c_6 + 1.0/c_7 + 0.9/c_8
\end{aligned}
$$

$$
\begin{aligned}
G^{10} = {} & 0.0/s_1 + 0.0/s_2 +\\
& +0.0/s_3 + 0.1/s_4 + 0.1/s_5 + 0.2/s_6 + 0.2/s_7 + 0.3/s_8 + 0.4/s_9 + 1.0/s_{10}
\end{aligned}
$$

We assume that the main parameters are: the population size is 250, the number of trials is 1,000, the crossover rate is 0.6, and the mutation rate is 0.001.

The best ("optimal") result obtained is

$$
\begin{array}{l}
u_0 = c_1 \; u_1 = c_0 \; u_2 = c_0\\
u_3 = c_1 \; u_4 = c_0 \; u_5 = c_1\\
u_6 = c_7 \; u_7 = c_7 \; u_8 = c_7\\
u_9 = c_6
\end{array}
$$

with $\mu_D(. \mid .) = 0.5553$.

A similar best result was obtained by using the BEA:

$$
\begin{array}{l}
u_0 = c_1 \; u_1 = c_0 \; u_2 = c_0\\
u_3 = c_1 \; u_4 = c_0 \; u_5 = c_1\\
u_6 = c_7 \; u_7 = c_7 \; u_8 = c_7\\
u_9 = c_6
\end{array}
$$

with $\mu_D(. \mid .) = 0.5554$.

The results are quite similar, which is no surprise for such a small example, and the speed of convergence is also similar, though – roughly speaking, a very good value of $\mu_D(. \mid .)$ was obtained after ca. 500 iterations in the case of GA, and after ca. 400 in the case of BEA. However, since a further extension of the BEA, the so called *bacterial memetic algorithms* (BMA) have been reported by Kóczy and his collaborators [2,3,4]) to have a good efficiency, mainly due to the use of an inside gradient based local optimization, we believe that the BEA can be an interesting solution for our problem, and a point of departure for a further extension to the BMA so that is why we first proposed the use of the BEA.

6 Concluding Remarks

We proposed the use of a bacterial evolutionary algorithm (BEA) for the multistage fuzzy control of a stochastic system. The algorithm yielded good results, better than the traditional genetic algorithm(GA). However, as our approach is directly extendable to accomodate the bacterial memetic algorithm (BMA) proposed by Kóczy et al. [2,3,4], which was reported to have shown an even better efficiency, we think that our approach is relevant by being a step in a proper direction.

References

1. Bellman, R.E., Zadeh, L.A.: Decision making in a fuzzy environment. Management Science 17, 141–164 (1970)
2. Botzheim, J., Cabrita, C., Kóczy, L.T., Ruano, A.: Genetic and bacterial programming for B-spline neural networks design. Journal of Advanced Computational Intelligence and Intelligent Informatics 11(2), 220–231 (2007)
3. Botzheim, J., Cabrita, C., Kóczy, L.T., Ruano, A.: Fuzzy rule extraction by bacterial memetic algorithms. International Journal of Intelligent Systems 24(3), 312–339 (2009)
4. Gál, L., Kóczy, L.T.: Advanced bacterial memetic algorithms. Acta Technica Jauriniensis, Series Intelligentia Combinatorica 1(3), 481–498 (2008)
5. Herrera, F., Lozano, M., Verdegay, J.L.: Tackling real-coded genetic algorithms: Operators and tools for behavioural analysis. Artificial Intelligence Review 12(4), 265–319 (2008)
6. Kacprzyk, J.: Multistage Decision Making under Fuzziness, Verlag TÜV Rheinland, Cologne (1983)
7. Kacprzyk, J.: Stochastic systems in fuzzy environments: control. In: Singh, M.G. (ed.) Systems and Control Encyclopedia, pp. 4657–4661. Pergamon Press, Oxford (1987)
8. Kacprzyk, J.: Multistage control under fuzziness using genetic algorithms. Control and Cybernetics 25, 1181–1215 (1996)
9. Kacprzyk, J.: Multistage Fuzzy Control. Wiley, Chichester (1997)
10. Kacprzyk, J.: Multistage control of a stochastic system under fuzzy goals and constraints using a genetic algorithm. In: Proceedings of IFSA 1997 – Seventh International Fuzzy Systems Association World Congress, Prague, Czech Rep., vol. II, pp. 306–311 (1997)

11. Kacprzyk, J.: A genetic algorithm for the multistage control of a fuzzy system in a fuzzy environment. Mathware and Soft Computing I(3) 219–232 (1997)
12. Kacprzyk, J.: Multistage control of a stochastic system in a fuzzy environment using a genetic algorithm. International Journal of Intelligent Systems 13, 1011–1023 (1998)
13. Kacprzyk, J.: Fuzzy dynamic programming: interpolative reasoning for an efficient derivation of optimal control policies. Control and Cybernetics 42(1), 63–84 (2013)
14. Kacprzyk, J., Staniewski, P.: A new approach to the control of stochastic systems in a fuzzy environment. Archiwum Automatyki i Telemechaniki XXV, 433–443 (1980)
15. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer, Heidelberg (1996)
16. Michalewicz, Z., Janikow, C.: Genetic algorithms for numerical optimization. Statistics and Computing 1, 75–91 (1991)
17. Nawa, N.E., Furuhashi, T., Hashiyama, T., Uchikawa, Y.: A Study on the discovery of relevant fuzzy rules using pseudo-bacterial genetic algorithm. IEEE Trans. on Industrial Electronics 46(6), 1080–1089 (1999)
18. Nawa, N.E., Furuhashi, T.: Fuzzy system parameters discovery by bacterial evolutionary algorithm. IEEE Transactions on Fuzzy Systems 7(5), 608–616 (1999)

The Role of a T-norm and Partitioning in Fuzzy Association Analysis

Jiří Kupka and Pavel Rusnok

Centre of Excellence IT4Innovations, Division of the University of Ostrava
Institute for Research and Applications of Fuzzy Modeling
30. dubna 22, Ostrava, Czech Republic
{jiri.kupka,pavel.rusnok}@osu.cz

Abstract. Fuzzy association analysis extracts relationships from data. The result of fuzzy association analysis depends on a chosen t-norm that is used for calculating confidence and support measures of mined association rules. We show that the set of mined association rules might change depending on the t-norm. We measure the distances of sets of mined rules with different t-norms and also with set of rules mined by crisp association analysis. We experiment with various datasets and partitioning methods to examine relationships of mined rules by different t-norms. Our experiments shed new light on application of fuzzy association mining and confirm that fuzzy association analysis usually brings signifficantly different results when compared to results given by crisp (non-fuzzy) association analysis.

Keywords: fuzzy association analysis, t-norm, association rules.

1 Introduction

Association rule mining is well established field in data mining [1], originally studied in more general framework under the name GUHA [5]. The original methods were designed for boolean variables having only two values 0 and 1. The generalization to multinomial variables was straightforward as every category is translated into its own boolean variable. Quantitative variables are translated into nominal variables by a discretization of the interval from which the values are, but in this case a lot of information (e.g. distribution) is lost. Alternative approach to that is to define fuzzy sets on the variable domain (for definitions see Section 2).

Recently there was a discussion whether the presence of fuzzy association rule mining is defendable. In [9] Verlinde et. al compared the rules mined by fuzzy association analysis and crisp association analysis. They restricted themselves to a smaller amount of variables in the data studied, to rules with only one antecedent and one succedent mined and only one partitioning method, which is not sound towards fuzzy association mining – we show it in Section 3.1. Furthermore, only the order based on confidence or support measure of the rules were compared with Spearman correlation index. In this very special scenario the

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_33

fuzzy association analysis was denounced as not useful method for calculating support and confidence measures of association rules.

In response to [9] Hüllermeier and Yi in [6] explored association mining in a more general setting. Association rules with more than one antecedent were mined and various partitioning techniques were used. The differences between top-50 and top-100 rules sorted by confidence were studied and in case of antecedents of length 4 the similarity of mined rules by crisp association analysis and fuzzy association analysis disappeared absolutely.

In [6] only the minimum t-norm was used for calculating the supports of rules. In [9] also other t-norms were used but was claimed it does not make a difference. We consider all three prominent t-norms, i.e. Łukasiewicz, Minimum and Product t-norm, respectively. We also argue against fuzzy c-means as a technique for partitioning the domains of variables as it was used in [9]. We propose alternative data-driven partitioning technique to fuzzy c-means that is semantically sound to fuzzy association mining.

We present here neither the arguments for or against the fuzzy association analysis, but we empirically investigate the consequences of choosing fuzzy or crisp association analysis for mining information from quantitative variables. Moreover, in case of fuzzy association analysis, we discuss the influence of a chosen t-norm. The question is not which one is better, but whether we want to model the imprecisions of data or get rid of them before the search for a model starts and if we want to model the imprecision than how particular t-norm influences our data mining result.

The next section presents some preliminaries. In Section 3 we describe the partitionings of variable domains. Section 4 contains a brief summary of data sets used in our experiments. We define measures for describing the relationships of mined rules in Section 5. We follow with a discussion of our results in Section 6 and conclude with Section 7.

2 Preliminaries

In this section we are going to define an association rule between real-valued attributes in a data set. Let there be a 2 dimensional data set $\mathcal{D}$ with rows/objects $o_1, \ldots, o_n$ and columns/attributes $A_1, A_2, \ldots, A_m$. We will denote the value of attribute A_i for object o_j in data set $\mathcal{D}$ as $A_i(o_j)$. We define a domain of an attribute A_i as $\mathrm{dom}(A_i) = [\mathrm{min}_i, \mathrm{max}_i]$, where $\mathrm{min}_i = \min_j A_i(o_j)$ and $\mathrm{max}_i = \max_j A_i(o_j)$.

For every attribute A_i we define a set of *basic fuzzy attributes* $A_i^1, A_i^2, \ldots, A_i^{p_i}$ which are mappings from $\mathrm{dom}(A_i)$ to $[0,1]$. They are in fact fuzzy sets defined on domains of attributes. We will simply write $A_i^j(o_k)$ instead of $A_i^j(A_i(o_k))$ to denote the value of a basic fuzzy attribute A_i^j for an object o_k (i.e. membership degree of o_k in a fuzzy set A_i^j).

We define *fuzzy attributes* as combination (resp. a conjunction) of basic fuzzy attributes by a t-norm $\otimes$

$$A(o_k) = \bigotimes_i A_i^j(o_k).$$

Basic t-norms we are going to use are the Łukasiewicz, Minimum and Product t-norm, respectively, denoted (ł,m,p):

$$\text{ł}(a,b) = \max(a+b-1,0), \qquad m(a,b) = \min(a,b), \qquad p(a,b) = a \cdot b.$$

Let A and B be disjoint and non-empty sets of fuzzy attributes then we can define a *support* ($\mathrm{supp}_\otimes(A \to B)$) and *confidence* ($\mathrm{conf}_\otimes(A \to B)$) of a rule $(A \to B)$ in the following way:

$$\mathrm{supp}_\otimes(A \to B) = \frac{\sum_{o \in \mathcal{D}} A(o) \otimes B(o)}{n}, \tag{1}$$

$$\mathrm{conf}_\otimes(A \to B) = \frac{\sum_{o \in \mathcal{D}} A(o) \otimes B(o)}{\sum_{o \in \mathcal{D}} A(o)}. \tag{2}$$

We call A in a rule $(A \to B)$ *antecedent* and B *succedent.* If A (resp. B) is a fuzzy attribute combined from n basic fuzzy attributes then we say that a rule $(A \to B)$ has n antecedents (resp. n succedents). In our experiments, we searched for rules with only one succedent.

Usually a t-norm chosen in applications of fuzzy association mining is the Minimum t-norm. And there are also semantical reasons for that as it was shown in [3] that it is the only t-norm for which the following holds: if for each $o \in \mathcal{D}$ holds $A(o) \leq B(o)$ then equality $\mathrm{conf}_\otimes(A \to B) = 1$ is true. In [3] general class of admissible t-norms for calculating fuzzy association rules is defined and showed to fulfill the condition: $\forall a, b, c \in [0,1] : (a \leq b) \Rightarrow (b \otimes c) - (a \otimes c) \leq b - a$. The Łukasiewicz t-norm being the smallest of them – originally shown in [7].

The set of all association rules mined from data $\mathcal{D}$ using a t-norm $\otimes$ is denoted as $\mathcal{R}_\otimes$. Usually in the output of an association analysis the rules with highest confidence are given. We denote the first n rules with highest confidence mined from data $\mathcal{D}$ using a t-norm $\otimes$ as $\mathcal{R}^n_\otimes$.

By *crisp* associations we mean association rules mined from data where attributes have only values 0 and 1. The definition of support and confidence of crisp association rule is the same as for fuzzy association rule defined by (1) and (2) independently on the chosen t-norm. It is easy to see that our definition of support and confidence coincides with the definitions in classical literature on association rule mining.

In our experiments we define crisp attributes via partioning the domains of attributes into intervals. Based on these intervals we define fuzzy attributes (for details see Section 3). Hence there is the same amount of crisp and fuzzy attributes in paralell. The crisp association rules mined from data $\mathcal{D}$ (paralell crisp attributes) are denoted $\mathcal{R}_c$.

3 Partitioning

Verlinde et. al in [9] partiotioned every variable through clustering into three clusters with fuzzy c-means, corresponding to linguistic terms *Small*, *Medium* and *Big*. The corresponding cluster centers were taken as the initial centers for crisp k-means clustering. The resulting crisp clusters were more or less the same as applying the maximum function to the membership degrees of fuzzy clusters. Nonetheless, the presence of outliers lead to unwanted membership functions. In Figure 1, we can see that highest values are Big to the degree 0.5 and are at the same time Small to the degree 0.25, which does not make sense. Even when we get rid of the outliers by preprocessing the issue of non-monotonicity remains.

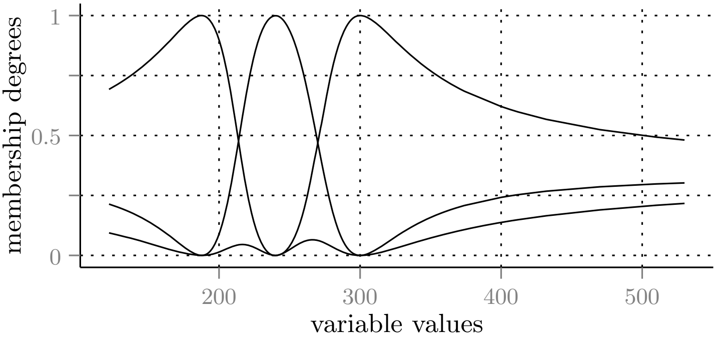


Fig. 1. Effect of outliers on membership functions

In our experiments we use three crisp partitions (namely, clustering-, equi-width- and equi-frequency-based one) and three fuzzy ones (each of them induced from a chosen crisp partition). We cluster the data with k-means algorithm. The clusters centers are assigned degree 1 to the respective fuzzy sets and 0 to neighbouring clusters. The inbetween clusters means are assigned a degree 0.5 to both clusters and the points are linearly connected. You can see the resulting partitions in Figure 2. Our data driven partitioning has more correct semantical base and the functions are non-decreasing (resp. non-increasing) for terms *Big* (resp. *Small*). We call this partition *cluster partition.*

We also used the equi-width and equi-frequency methods of crisp partitioning. We created the membership functions parallel to the crisp intervals resulting from equi-width (resp. equi-frequency) method analogically to the clustering case.

K-means is unstable and prone to get stuck in local minima and therefore we performed multiple runs in our experiments to obtain optimal solution.

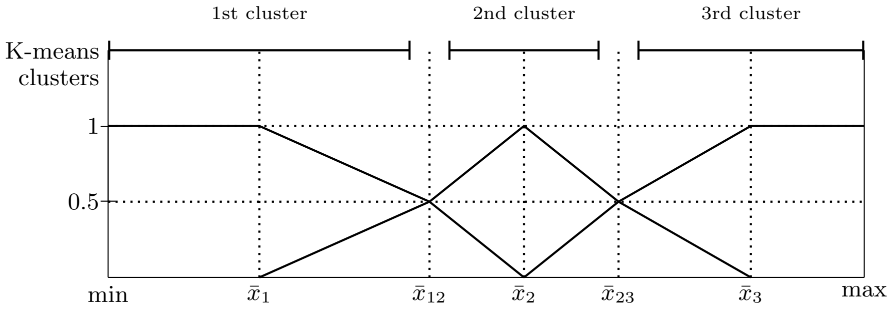


Fig. 2. Triangular partition derived from k-means clustering. Where $\bar{x}_i$ is the mean of i-th cluster and $\bar{x}_{ij}$ is the mean of the i-th cluster maximum and the j-th cluster minimum.

3.1 Semantic Issues in Comparing the Crisp and Fuzzy Association Mining

In [6] and [9] the sets of rules mined by fuzzy association analysis and crisp association analysis were compared. This article also provides this comparison partly while paying attention more to comparing different sets of mined rules depending on the chosen t-norm. We want to mention one semantical issue that is hidden in comparing the fuzzy and crisp associations. As it was stated in Section 2, crisp associations are special case of fuzzy associations, but to compare fuzzy associations with crisp we define the same amount of crisp/fuzzy attributes. It is possible to view the crisp attributes as fuzzy attributes given by characteristic functions. Comparison of $\mathcal{R}_\otimes$ with $\mathcal{R}_c$ is in fact comparison of two results of fuzzy association analysis but in the latter case with different fuzzy attributes. We are basically comparing two absolutely different sets of rules. We should bear in mind this semantical issue when interpreting our results. However, to somehow compare fuzzy and crisp association analysis we have to do such semantical skip.

4 Data Sets

We have used four different data sets for our experiments. We have purposefully chosen differing types of datasets to eliminate its influence on our findings. The first data set `Entry`, which was also used in [6] and [9] consists of medical data about patients.[1] Second data set `Abalone` consists of physical measurements for predicting the age of abalone. The third dataset `SML2010` consists of time series collected from a monitoring system mounted in a domotic house. The last dataset `Yeast` consists of various scores for predicting the cellular localization

[1] This data set was obtained from website
`http://lisp.vse.cz/challenge/ecmlpkdd2004`

sites in proteins. The latter three data sets were downloaded from URI Machine learning repository [2]. On every data set, the results of our experiments were the same in relation to the trends of distances and we present in Tables 2-10 only the results calculated on data set `Yeast`.

5 Measures for Comparison of Mined Rules Set

Before we start discussing our results we want to mention one peculiar feature of association rules. Assume you are given data $\mathcal{D}$ and mapped all the attributes in data $\mathcal{D}$ into fuzzy attributes. Then in general for fuzzy attributes A_i, A_j, B_i and B_j, t-norms $\otimes_1$, $\otimes_2$ from inequality $\mathrm{conf}_{\otimes_1}(A_i \to B_i) < \mathrm{conf}_{\otimes_1}(A_j \to B_j)$ does not follow that $\mathrm{conf}_{\otimes_2}(A_i \to B_i) < \mathrm{conf}_{\otimes_2}(A_j \to B_j)$.

Example 1. For example look at the simple data example from Table 1. By calculation of confidence according to (2) with values from Table 1 we obtain the following relations between confidences: $\mathrm{conf}_{ł}(A_1 \to B_1) < \mathrm{conf}_{ł}(A_2 \to B_2)$ but we get $\mathrm{conf}_m(A_1 \to B_1) > \mathrm{conf}_m(A_2 \to B_2)$. For another pair of rules the situation is opposite: $\mathrm{conf}_{ł}(A_3 \to B_3) > \mathrm{conf}_{ł}(A_4 \to B_4)$ but we get $\mathrm{conf}_m(A_3 \to B_3) < \mathrm{conf}_m(A_4 \to B_4)$. The product t-norm orders the first two rules according to confidence in the same way as the Łukasiewicz t-norm and in the second case as the Minimum t-norm.

Table 1. Example of different rules ordering depending on chosen t-norm

	A_1	B_1	A_2	B_2	A_3	B_3	A_4	B_4
o_1	0.8	0.8	0.8	0.8	0.8	0.2	0.6	0.8
o_2	0.8	0.8	1	0.9	0.9	0.9	0.5	0.5

We define for a rule $r \in \mathcal{R}^n_\otimes$ its rank as $rank(r) = j$ where j stands for the j-th best position in $\mathcal{R}^n_\otimes$ when $\mathcal{R}^n_\otimes$ is ordered descending according to the confidence measure. K_n was used in [6] and originally proposed in [4].

For two sets $\mathcal{R}^n_{\otimes_1}$ and $\mathcal{R}^n_{\otimes_2}$, where $\otimes_i \in \{ł, p, m\}$ ordered by confidence we define a distance $K_n(\mathcal{R}^n_{\otimes_1}, \mathcal{R}^n_{\otimes_2}) \in [0, 1]$ as:

$$K_n(\mathcal{R}^n_{\otimes_1}, \mathcal{R}^n_{\otimes_2}) = \frac{1}{n \cdot (n+1)} \sum_{r \in \mathcal{R}^n_{\otimes_1} \cup \mathcal{R}^n_{\otimes_2}} |rank_1(r) - rank_2(r)|, \tag{3}$$

where $rank_i(r)$ is the rank of a rule r in $\mathcal{R}^n_{\otimes_i}$. If $r \notin \mathcal{R}^n_{\otimes_i}$ then $rank_i(r) = n + 1$ $(i \in \{1, 2\})$. This definition is extended also to the case of $\mathcal{R}^n_c$.

We also define absolute difference D_n of two rule sets as a cardinality of a set difference:

$$D_n(\mathcal{R}^n_{\otimes_1}, \mathcal{R}^n_{\otimes_2}) = \#\{\mathcal{R}^n_{\otimes_1} \backslash \mathcal{R}^n_{\otimes_2}\}. \tag{4}$$

$K_n = 0$ when the rules in $\mathcal{R}^n_{\otimes_1}$ and $\mathcal{R}^n_{\otimes_2}$ are the same and ordered the same way. $K_n = 1$ when there is no rule in common ($\mathcal{R}^n_{\otimes_1} \cap \mathcal{R}^n_{\otimes_2} = \emptyset$). $D_n \in [0, n]$ and might equal 0 even when K_n is non zero. Also note that $D_n = n$ when $K_n = 1$.

6 Results

The distances of all mined sets from Yeast data set are in Tables 2-10. Generally with increasing number of antecedents the sets grow appart, but there are interesting patterns showing up.

Rules mined with the Łukasiewicz t-norm tend to be closer to crisp association rules than the rules mined by other t-norms in case of equi-width partitioning. In equi-width partitioning we are more likely to encounter the bordering effect described e.g. in [8]. The Łukasiewicz t-norm reduces the bordering effect as values near 0.5 are mapped near 0, see Tables 8-10. In this context we might view Łukasiewicz t-norm as least fuzzy.

For rule sets with only one antecedent and cluster partitioning (Table 2) we obtain results similar to Verlinde et. al in [9]. This is however not the case for equi-frequency partitioning, see Table 5.

As stated, increasing the number of antecedents increases the distances, but not in the same extent for minimum and product t-norm that tend to be still close enough (see Tables 2-4, 5-7 or 8-10).

It was shown in [6] that the distance K_{100} between $\mathcal{R}^{100}_c$ and $\mathcal{R}^{100}_m$, reaches 1 when considering rules with 4 antecedents. In Table 10, we can see that the distance between $\mathcal{R}^{200}_t$ and $\mathcal{R}^{200}_p$ is 1 already for rules with 3 antecedents. We may obtain more differing results with fuzzy association analysis when only considering different t-norms then when switching between fuzzy and crisp case. This we consider as the most interesting result of our experiments.

Table 2. Distances of rules with 1 antecedent mined from dataset Yeast with cluster partition

$D_{200} \backslash K_{200}$	$\mathcal{R}^{200}_c$	$\mathcal{R}^{200}_t$	$\mathcal{R}^{200}_m$	$\mathcal{R}^{200}_p$
$\mathcal{R}^{200}_c$	x	0.09	0.12	0.11
$\mathcal{R}^{200}_t$	12	x	0.12	0.08
$\mathcal{R}^{200}_m$	16	17	x	0.04
$\mathcal{R}^{200}_p$	15	12	6	x

Table 3. Distances of rules with 2 antecedents mined from dataset Yeast with cluster partition

$D_{200} \backslash K_{200}$	$\mathcal{R}^{200}_c$	$\mathcal{R}^{200}_t$	$\mathcal{R}^{200}_m$	$\mathcal{R}^{200}_p$
$\mathcal{R}^{200}_c$	x	0.4	0.38	0.34
$\mathcal{R}^{200}_t$	68	x	0.53	0.48
$\mathcal{R}^{200}_m$	62	84	x	0.11
$\mathcal{R}^{200}_p$	54	74	14	x

Table 4. Distances of rules with 3 antecedents mined from dataset `Yeast` with cluster partition

$D_{200}\backslash K_{200}$	$\mathcal{R}_c^{200}$	$\mathcal{R}_t^{200}$	$\mathcal{R}_m^{200}$	$\mathcal{R}_p^{200}$
$\mathcal{R}_c^{200}$	x	0.85	0.83	0.89
$\mathcal{R}_t^{200}$	161	x	0.95	0.96
$\mathcal{R}_m^{200}$	145	184	x	0.22
$\mathcal{R}_p^{200}$	163	190	45	x

Table 5. Distances of rules with 1 antecedent mined from dataset `Yeast` with equi-frequency partition

$D_{200}\backslash K_{200}$	$\mathcal{R}_c^{200}$	$\mathcal{R}_t^{200}$	$\mathcal{R}_m^{200}$	$\mathcal{R}_p^{200}$
$\mathcal{R}_c^{200}$	x	0.42	0.41	0.41
$\mathcal{R}_t^{200}$	48	x	0.19	0.09
$\mathcal{R}_m^{200}$	47	18	x	0.11
$\mathcal{R}_p^{200}$	47	13	8	x

Table 6. Distances of rules with 2 antecedents mined from dataset `Yeast` with equi-frequency partition

$D_{200}\backslash K_{200}$	$\mathcal{R}_c^{200}$	$\mathcal{R}_t^{200}$	$\mathcal{R}_m^{200}$	$\mathcal{R}_p^{200}$
$\mathcal{R}_c^{200}$	x	0.77	0.66	0.54
$\mathcal{R}_t^{200}$	149	x	0.7	0.62
$\mathcal{R}_m^{200}$	129	122	x	0.39
$\mathcal{R}_p^{200}$	102	101	67	x

Table 7. Distances of rules with 3 antecedents mined from dataset `Yeast` with equi-frequency partition

$D_{200}\backslash K_{200}$	$\mathcal{R}_c^{200}$	$\mathcal{R}_t^{200}$	$\mathcal{R}_m^{200}$	$\mathcal{R}_p^{200}$
$\mathcal{R}_c^{200}$	x	0.93	0.62	0.53
$\mathcal{R}_t^{200}$	179	x	0.84	0.91
$\mathcal{R}_m^{200}$	113	153	x	0.51
$\mathcal{R}_p^{200}$	100	172	93	x

Table 8. Distances of rules with 1 antecedent mined from dataset `Yeast` with equi-width partition

$D_{200}\backslash K_{200}$	$\mathcal{R}_c^{200}$	$\mathcal{R}_t^{200}$	$\mathcal{R}_m^{200}$	$\mathcal{R}_p^{200}$
$\mathcal{R}_c^{200}$	x	0.16	0.19	0.18
$\mathcal{R}_t^{200}$	22	x	0.15	0.11
$\mathcal{R}_m^{200}$	29	24	x	0.06
$\mathcal{R}_p^{200}$	25	12	14	x

Table 9. Distances of rules with 2 antecedents mined from dataset `Yeast` with equi-width partition

$D_{200}\backslash K_{200}$	$\mathcal{R}_c^{200}$	$\mathcal{R}_t^{200}$	$\mathcal{R}_m^{200}$	$\mathcal{R}_p^{200}$
$\mathcal{R}_c^{200}$	x	0.38	0.62	0.76
$\mathcal{R}_t^{200}$	53	x	0.59	0.7
$\mathcal{R}_m^{200}$	112	94	x	0.21
$\mathcal{R}_p^{200}$	141	123	35	x

Table 10. Distances of rules with 3 antecedents mined from dataset `Yeast` with equi-width partition

$D_{200}\backslash K_{200}$	$\mathcal{R}_c^{200}$	$\mathcal{R}_t^{200}$	$\mathcal{R}_m^{200}$	$\mathcal{R}_p^{200}$
$\mathcal{R}_c^{200}$	x	0.59	0.85	0.99
$\mathcal{R}_t^{200}$	91	x	0.89	1
$\mathcal{R}_m^{200}$	166	171	x	0.43
$\mathcal{R}_p^{200}$	191	197	68	x

7 Conclusions and Future Work

Our results show that it is reasonable to mine fuzzy associations with various t-norms, because of different results that might be obtained. In our future work we are going to device some techniques that would choose the best from various data mining runs or valid rules for all t-norms.

Using k-means is not optimal and application of clustering that reflects shapes of data sets should be considered. Other possibility for improvement is to extend the study to other confidence measures (resp. implicational quantifiers [5]). A study of distinct shapes of partitions will be also included in our future work.

Acknowledgments. This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). Furthermore, we gratefully acknowledge partial support of project SGS13/PřF/2014 of the University of Ostrava.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
2. Bache, K., Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
3. Dubois, D., Hüllermeier, E., Prade, H.: A systematic approach to the assessment of fuzzy association rules. Data Mining and Knowledge Discovery 13(2), 167–192 (2006)
4. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM Journal on Discrete Mathematics 17(1), 134–160 (2003)
5. Hájek, P., Havránek, T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory). Springer-Verlag (1978)
6. Hullermeier, E., Yi, Y.: In defense of fuzzy association analysis. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 37(4), 1039–1043 (2007)
7. Schweizer, B., Sklar, A.: Probabilistic Metric Spaces. North-Holland, New York (1983)
8. Sudkamp, T.: Examples, counterexamples, and measuring fuzzy associations. Fuzzy Sets and Systems 149(1), 57–71 (2005)
9. Verlinde, H., De Cock, M., Boute, R.: Fuzzy versus quantitative association rules: a fair data-driven comparison. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 36(3), 679–684 (2005)

Author Index