STUDIES IN FUZZINESS AND SOFT COMPUTING

Christian Borgelt
María Ángeles Gil
João M.C. Sousa
Michel Verleysen (Eds.)

Towards Advanced Data Analysis by Combining Soft Computing and Statistics

Springer

Studies in Fuzziness and Soft Computing 285

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/2941

Christian Borgelt, María Ángeles Gil,
João M.C. Sousa, and Michel Verleysen (Eds.)

Towards Advanced Data Analysis by Combining Soft Computing and Statistics

Springer

Editors
Dr. Christian Borgelt
European Centre for Soft Computing
Mieres
Spain

María Ángeles Gil
Universidad de Oviedo
Spain

João M.C. Sousa
Technical University Lisbon
Portugal

Michel Verleysen
Université Catholique de Louvain (UCL)
Louvain-la-Neuve
Belgium

ISSN 1434-9922
ISBN 978-3-642-30277-0
DOI 10.1007/978-3-642-30278-7
Springer Heidelberg New York Dordrecht London

e-ISSN 1860-0808
e-ISBN 978-3-642-30278-7

Library of Congress Control Number: 2012939124

Printed on acid-free paper

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

Preface

This book is the Final Action Publication of COST Action IC0702: "Combining Soft Computing Techniques and Statistical Methods to Improve Data Analysis Solutions" (SoftStat).

The main objective of this Action was to strengthen the dialog between the statistics and soft computing research communities in order to cross-pollinate both fields and generate mutual improvement activities.

Soft computing, as an engineering science, and statistics, as a branch of mathematics, emphasize different aspects of data analysis. Soft computing focuses on obtaining working solutions quickly, accepting approximations and unconventional approaches. Its strength lies in its flexibility to create models that suit the needs arising in applications (context of discovery, model generation). In addition, it emphasizes the need for intuitive and interpretable models, which are tolerant to imprecision and uncertainty.

Statistics is more rigorous and focuses on establishing objective conclusions based on experimental data by analyzing the possible situations and their (relative) likelihood (context of justification, model validation). It emphasizes the need for mathematical methods and tools to assess solutions and guarantee performance.

Bringing the two fields closer together enhances the robustness and generalizability of data analysis methods, while preserving the flexibility to solve real-world problems efficiently and intuitively.

This book contains 28 contributions from various members of COST Action IC0702, many of which are the outcome of short-term scientific missions (STSMs), that is, of visits that a (preferably early-stage) researcher paid to researcher in another country. In this way it is demonstrated that the Action made good use of this tool and nourished significant scientific collaborations that can be expected to continue beyond the end of the Action.

The financial support provided by the COST Office and the European Science Foundation (ESF) for both the Action as a whole and this Final Action Publication in particular is gratefully acknowledged.

March 2012

Christian Borgelt (Mieres, Spain)
María Ángeles Gil (Oviedo, Spain)
João M.C. Sousa (Lisbon, Portugal)
Michel Verleysen (Louvain, Belgium)

Contents

Arithmetic and Distance-Based Approach to the Statistical Analysis of Imprecisely Valued Data

Angela Blanco-Fernández[1], María Rosa Casals[1], Ana Colubi[1], Renato Coppi[3], Norberto Corral[1], Sara de la Rosa de Sáa[1], Pierpaolo D'Urso[3], Maria Brigida Ferraro[3], Marta García-Bárzana[1], María Ángeles Gil[1], Paolo Giordani[3], Gil González-Rodríguez[1], María Teresa López[1], María Asunción Lubiano[1], Manuel Montenegro[1], Takehiko Nakama[2], Ana Belén Ramos-Guajardo[1,2], Beatriz Sinova[1], and Wolfgang Trutschnig[2]

Abstract Most of the research developed in the last years by the SMIRE Research Group concerns the statistical analysis of imprecisely (set- and fuzzy set)-valued experimental data. This analysis has been based on an approach considering the usual arithmetic for these data as well as suitable metrics between them. The research perfectly fits into the research directions of the COST Action IC0702, which has been particularly helpful for scientific activities, discussions and exchanges associated with group members. The main objective of this paper is to summarize some of the main recent advances of the SMIRE Research Group.

1 Introduction

Traditionally, Statistical Data Analysis assumes that experimental data are either quantitative or qualitative. Most of the statistical techniques have been developed to deal with quantitative data. They allow us to exploit the wealth of information concerning the diversity and variability of these data.

Nevertheless, qualitative experimental data are also present in real-life situations, and there is a rather limited class of statistical procedures to analyze these data. Furthermore, most of these procedures do not exploit all the information contained in the data.

In this respect, when we consider random experiments in which the attribute to be 'measured' is a range, a fluctuation, a grouping, etc., instead of using a qualitative scale we can employ the scales of interval or set val-

[1] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, E-33007 Oviedo, Spain, smire@uniovi.es
[2] European Centre for Soft Computing, E-33600 Mieres, Spain
[3] Sapienza Università di Roma, I-00185 Roma, Italy

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 1–18.
springerlink.com

ues. On the other hand, when we consider random experiments in which the attribute to be 'measured' is intrinsically imprecise but gradual (see Dubois and Prade [20]), as often happens with ratings, valuations, etc, then instead of using a qualitative scale we can employ the expressive and flexible scale of fuzzy numbers or values.

Although set- and fuzzy set-valued data cannot be statistically analyzed as easily as real/vectorial valued data, many methods for this purpose have been developed in the last years. In general, these methods exploit all the information available in the analyzed data, so that the diversity and variability (due to the randomness associated with the generating data mechanisms) is not lost, and the imprecision (due to data's nature) is well-captured.

Two key tools in formalizing and developing the problems, models and techniques to analyze set/fuzzy set-valued experimental data are: the choice of suitable metrics between data, and the choice of an appropriate model for the random mechanism generating them.

The latter model will be that of set/fuzzy set-valued random elements (REs, for short), i.e., Borel-measurable mappings w.r.t. topologies associated with suitable metrics on the spaces of set/fuzzy set values, leading to the concepts of random compact convex set (RS, for short) and random fuzzy set (RFS, for short). It should be emphasized and clarified that the choice of these notions were intended to model random mechanisms generating 'existing' set/fuzzy set-valued data. Coherently, the statistical methods handling data obtained from REs aim to draw statistical conclusions on these elements.

In this paper RSs and RFSs are treated from an 'ontic' perspective (see Dubois and Prade [20]). From a theoretical point of view the related techniques could be applied to the 'epistemic' perspective, in accordance with which data are imprecise perceptions of existing real/vectorial-valued data which cannot be measured exactly. However, RSs and RFSs as considered in this paper, although formally applicable, would not really capture the essence of the situation to be modeled and the corresponding statistical techniques wouldn't allow practitioners to draw conclusions on the relevant random element: the underlying unobservable random variable/vector. From the "ontic" perspective the expression "imprecise data" (and "imprecisely-valued" random elements) is used in opposition to "precise data", to express that the considered experimental values cannot be described by a single (precise) real number or vector.

In previous papers (those published before the COST Action IC0702 started), the SMIRE RG presented statistical developments concerning: the 'point' estimation of some parameters of the distribution of an RE (like the population mean, for which 'point' estimation should be understood as set/fuzzy set-valued one, the population Fréchet-type variance, the population inequality, and so on); the one- and two-sample cases of testing hypotheses about means of REs; some rather introductory studies on other problems.

This paper aims to present a summary of several recent developments (in fact, initiated after the COST Action IC0702 started) of the SMIRE RG on

the topic of the Statistical Analysis of set/fuzzy set-valued experimental data. Details and discussions concerning the particular approaches can be found in the referred works. Although other, more or less related, problems have been investigated along the last years by the RG, we will focus on the following ones: in Section 2, we will recall some basic tools concerning metrics between imprecisely-valued data and simulations of these data; in Section 3, the developments on statistical methods to analyze the means (multi-sample case), and Fréchet-type variances for general REs will be mentioned, and a recent approach to define a robust centrality measure for an RE will be described for the one-dimensional case; in Section 4, some approaches will be discussed in connection with the regression analysis between data by assuming that either all of them or just the response ones are imprecisely-valued; in Section 5, an overview on clustering/classification methods to classify imprecisely-valued data is given. Finally, in Section 6 comments will be given on SAFD (Statistical Analysis of Fuzzy Data), an R package for statistical analysis of one dimensional fuzzy data.

2 Summary of Recent Results on Metrics and Simulations Concerning Imprecisely-valued Data

Metrics between imprecise (either set or fuzzy set)-valued data play a crucial role in their statistical analysis, along with the elementary arithmetic operations between these data. To properly recall some recent metrics, we first present some preliminary concepts, notations and results.

Let $\mathcal{K}_c(\mathbb{R}^p)$ (with $p \in \mathbb{N}$) be the space of nonempty compact convex subsets of $\mathbb{R}^p$. $\mathcal{K}_c(\mathbb{R})$ denotes the space of closed and bounded nonempty intervals.

The most usual and natural arithmetic on $\mathcal{K}_c(\mathbb{R}^p)$ is the one extending (through the image) the classical crisp operations. Thus, the two elementary operations, sum and product by a scalar, are defined so that for $K, K' \in \mathcal{K}_c(\mathbb{R}^p)$ and $\gamma \in \mathbb{R}$

$$K + K' = \{k + k' : k \in K,\, k' \in K'\}, \quad \gamma \cdot K = \{\gamma\, k : k \in K\}.$$

It is well-known that $(\mathcal{K}_c(\mathbb{R}^p), +, \cdot)$ does not have a linear but only a semilinear structure, since $K + (-1) \cdot K \neq \{\mathbf{0}\}$ (neutral element for $+$) unless K reduces to a singleton (in other words, K is 'degenerated' at a real/vectorial value).

Let $\mathcal{F}_c^*(\mathbb{R}^p)$ (with $p \in \mathbb{N}$) be the space of fuzzy set values in $\mathbb{R}^p$ such that any $\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R}^p)$ satisfies that for each $\alpha \in [0,1]$ the α-level set $\widetilde{U}_\alpha \in \mathcal{K}_c(\mathbb{R}^p)$ (where $\widetilde{U}_\alpha = \{x \in \mathbb{R}^p : \widetilde{U}(x) \geq \alpha\}$ if $\alpha > 0$, and $\widetilde{U}_0 = \mathrm{cl}\{x \in \mathbb{R}^p : \widetilde{U}(x) > 0\}$). $\mathcal{F}_c^*(\mathbb{R})$ denotes the space of the so-called (bounded) fuzzy numbers.

The most usual arithmetic on $\mathcal{F}_c^*(\mathbb{R}^p)$ is the one based on Zadeh's extension principle [64], which is levelwise equivalent to the usual set-valued arithmetic,

that is, for each $\alpha \in [0,1]$ if $\widetilde{U}, \widetilde{U}' \in \mathcal{F}_c^*(\mathbb{R}^p)$ and $\gamma \in \mathbb{R}$

$$(\widetilde{U} + \widetilde{U}')_\alpha = \widetilde{U}_\alpha + \widetilde{U}'_\alpha, \quad (\gamma \cdot \widetilde{U})_\alpha = \gamma \cdot \widetilde{U}_\alpha.$$

Also $(\mathcal{F}_c^*(\mathbb{R}^p), +, \cdot)$ does not have a linear but only a semilinear structure, since $\widetilde{U} + (-1) \cdot \widetilde{U} \neq \mathbf{1}_{\{\mathbf{0}\}}$ (neutral element for $+$) unless $\widetilde{U}$ reduces to the indicator function of a singleton (in other words, $\widetilde{U}$ is 'degenerated' at a real/vectorial value).

For the one-dimensional case many different (mostly L^2 and some L^1) distances can be found in the literature (see, for instance, [2], [18], [40], [62]). A metric which has been considered in several developments was introduced by Yang and Ko [63] to deal with LR fuzzy numbers $\widetilde{U} \in \mathcal{F}_{LR}(\mathbb{R}) \subset \mathcal{F}_c^*(\mathbb{R})$, denoted by $\widetilde{U} = (\widetilde{U}^m, \widetilde{U}^l, \widetilde{U}^r)$ with

$$\widetilde{U}(x) = \begin{cases} L\left(\frac{\widetilde{U}^m - x}{\widetilde{U}^l}\right) & \text{if } x \leq \widetilde{U}^m,\ \widetilde{U}^l > 0 \\ \mathbf{1}_{\{\widetilde{U}^m\}}(x) & \text{if } x \leq \widetilde{U}^m,\ \widetilde{U}^l = 0 \\ R\left(\frac{x - \widetilde{U}^m}{\widetilde{U}^r}\right) & \text{if } x > \widetilde{U}^m,\ \widetilde{U}^r > 0 \\ 0 & \text{if } x > \widetilde{U}^m,\ \widetilde{U}^r = 0 \end{cases}$$

where $\widetilde{U}^m \in \widetilde{U}_1$, $\widetilde{U}^l = \widetilde{U}^m - \inf \widetilde{U}_0$, $\widetilde{U}^r = \sup \widetilde{U}_0 - \widetilde{U}^m$, and $L, R : \mathbb{R} \to [0,1]$ are fixed convex upper semicontinuous functions so that $L(0) = R(0) = 1$, $L(z) = R(z) = 0$ for all $z \in \mathbb{R} \setminus [0,1]$ (see Dubois & Prade, 1978). This metric has been 'generalized' in $\mathbb{R}^3$ by Ferraro *et al.* [29] and has been widely applied afterwards papers as we will indicate later.

If $\widetilde{U}, \widetilde{U}' \in \mathcal{F}_{LR}(\mathbb{R})$, then Yang and Ko's metric is given by

$$D_{\mathfrak{l}\mathfrak{r}}(\widetilde{U}, \widetilde{U}') = \sqrt{\left[\widetilde{U}^m - \widetilde{U}'^m\right]^2 + \left[{}^{[\mathfrak{l}]}\widetilde{U} - {}^{[\mathfrak{l}]}\widetilde{U}'\right]^2 + \left[\widetilde{U}^{[\mathfrak{r}]} - \widetilde{U}'^{[\mathfrak{r}]}\right]^2},$$

where ${}^{[\mathfrak{l}]}\widetilde{U} = (\widetilde{U}^m - \mathfrak{l}\widetilde{U}^l)$, $\widetilde{U}^{[\mathfrak{r}]} = (\widetilde{U}^m + \mathfrak{r}\widetilde{U}^r)$, and $\mathfrak{l}, \mathfrak{r} \in (0, +\infty)$. In Yang and Ko [63], $\mathfrak{l}$ and $\mathfrak{r}$ have been usually chosen to be given by $\mathfrak{l} = \int_0^1 L^{-1}(\alpha)\, d\alpha$, $\mathfrak{r} = \int_0^1 R^{-1}(\alpha)\, d\alpha$ taking the shape of the involved LR fuzzy numbers into account. In particular, for $K, K' \in \mathcal{K}_c(\mathbb{R})$ if $\mathfrak{l} = \mathfrak{r}$ (which happens for Yang and Ko's usual choice) and $K^m = \operatorname{mid} K = (\inf K + \sup K)/2 =$ centre of K, then if we set $\operatorname{spr} K = (\sup K - \inf K)/2 =$ radius of K:

$$D_{\mathfrak{l}\mathfrak{l}}(K, K') = \sqrt{3[\operatorname{mid} K - \operatorname{mid} K']^2 + 2\mathfrak{l}^2[\operatorname{spr} K - \operatorname{spr} K']^2}.$$

In the p-dimensional case ($p \in \mathbb{N}$) several distances can be found in the literature (see, for instance, [18], [40], [41], [62]). Some new metrics have been introduced and discussed by Trutschnig *et al.* [60] and Trutschnig [59].

The first one is a multivariate version of Bertoluzza *et al.*'s metric [2] between intervals or between fuzzy numbers and is inspired on the equivalent

expression stated in the interval-valued case [33]. Let $\theta \in (0, +\infty)$ (often θ is supposed to be in $(0, 1]$) and let φ be an absolutely continuous probability measure on $([0, 1], \mathcal{B}_{[0,1]})$ with the mass function being positive in $(0, 1)$. Then, the metric by Trutschnig *et al.* assigns $\widetilde{U}, \widetilde{U}' \in \mathcal{F}_c^*(\mathbb{R}^p)$ the value

$$D_\theta^\varphi(\widetilde{U}, \widetilde{U}') = \sqrt{\int_{[0,1]} \left[d_\theta(\widetilde{U}_\alpha, \widetilde{U}'_\alpha)\right]^2 d\varphi(\alpha)}$$

$$= \sqrt{\int_{[0,1]} \int_{\mathbb{S}^{p-1}} \left[\operatorname{mid} \Pi\, \widetilde{U}_\alpha(u) - \operatorname{mid} \Pi\, \widetilde{U}'_\alpha(u)\right]^2 d\lambda_p(u)\, d\varphi(\alpha)}$$

$$\overline{+\,\theta \int_{[0,1]} \int_{\mathbb{S}^{p-1}} \left[\operatorname{spr} \Pi\, \widetilde{U}_\alpha(u) - \operatorname{spr} \Pi\, \widetilde{U}'_\alpha(u)\right]^2 d\lambda_p(u)\, d\varphi(\alpha)},$$

where $\mathbb{S}^{p-1}$ = unit sphere in $\mathbb{R}^p$, λ_p = normalized Lebesgue measure on $\mathbb{S}^{p-1}$, and $\Pi\, \widetilde{U}_\alpha(u)$ = projection of $\widetilde{U}_\alpha$ in the direction $u \in \mathbb{S}^{p-1}$.

The choice of θ allows us to weight the effect of the deviation in 'shape'/ 'imprecision' in contrast to the effect of the deviation in 'location', and φ enables to weight the importance of different α-levels. The metric above is intuitive, versatile and easy-to-handle and has been the one most frequently used in the statistical developments of the SMIRE RG.

The metric by Trutschnig [59] has been defined by considering certain distances between the sendographs of the fuzzy values. Several convergence and characterization results have been stated on the basis of this metric, although it has not yet been employed for statistical purposes.

Metrics between imprecisely-valued data will be relevant in quantifying errors in approximating some imprecise values by other ones, or to classify imprecise data in groups on the basis of their 'closeness'. They will also be useful to translate equalities between imprecise values into equalities of real numbers (more concretely, equalities of the distances between imprecise values to 0), which allows to overcome some of the drawbacks associated with the above-mentioned lack of linearity of the spaces $(\mathcal{K}_c(\mathbb{R}^p), +, \cdot)$ and $(\mathcal{F}_c^*(\mathbb{R}^p), +, \cdot)$.

The metrics above exhibit interesting properties, some topological equivalences, and they allow us to establish different Rådstrom-type isometries embedding the spaces $\mathcal{K}_c(\mathbb{R}^p)$ and $\mathcal{F}_c^*(\mathbb{R}^p)$ (actually, some wider ones) into cones of certain Hilbert spaces of functions or vectors (see Blanco-Fernández *et al.* [5], Ferraro *et al.* [29], and González-Rodríguez *et al.* [37], etc.).

The methodological developments to develop statistics with imprecisely-valued data have been based on the model for the random mechanisms generating imprecise data, namely, random sets and random fuzzy sets. The empirical developments have been based on the simulation of values from them.

Given a random experiment which is modeled by means of a probability space $(\Omega, \mathcal{A}, P)$, by a random compact convex set we mean (see, for instance, Puri and Ralescu [50], or Molchanov [46]) a mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R}^p)$ which is Borel-measurable w.r.t. $\mathcal{A}$ and the Borel σ-field generated by the topology induced by the well-known Hausdorff metric on $\mathcal{K}_c(\mathbb{R}^p)$.

Given a random experiment which is modeled by means of a probability space $(\Omega, \mathcal{A}, P)$, by a random fuzzy set we mean (see, Puri and Ralescu [51]) a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R}^p)$ (or even more general spaces) for which the α-level mappings $\mathcal{X}_\alpha : \Omega \to \mathcal{K}_c(\mathbb{R}^p)$ are random compact convex sets, whatever $\alpha \in [0, 1]$ may be. The original definition by Puri and Ralescu has been proved (see, for instance, Colubi *et al.* [10],[11], González-Rodríguez *et al.* [37]) to be equivalent to (or, at least, to imply) a certain Borel-measurability of the fuzzy set-valued mapping, which guarantees one can properly refer to notions like the independence of RFSs, or the (induced) distribution of an RFS. RFSs were originally coined by Puri and Ralescu as fuzzy random variables. Nevertheless, and to avoid misunderstanding with other concepts using the same name and to ease the interpretability of the notion we are using the term random fuzzy sets previously employed by some authors.

One of the main drawbacks in *simulating* RFSs is caused by the lack of realistic models for their distributions. In Colubi *et al.* [12] some ideas considering RFSs with values in some restrictive classes (like the class of trapezoidal fuzzy numbers in the one-dimensional fuzzy case, etc.) in combination with some classical real distributions have been outlined (say normal, chi-square, etc.). Recently, González-Rodríguez *et al.* [38] have suggested two different approaches following the usual ideas in Hilbert spaces. In this way, one of the suggested methods is based on the embedding allowing to identify fuzzy values with functional ones (see González-Rodríguez *et al.* [37]), the use of simulation techniques for Hilbert spaces and the posterior projection; some practical constraints are outlined. Another useful and easy-to-handle suggested method is the one making use of an extended orthonormal basis such that every fuzzy value can be approximated in terms of elements of the basis.

3 Summary of Recent Results on Inferences about Means and Fréchet's Variances, and on a More Robust Central Tendency Measure for Imprecisely-valued Data

Most of the statistical methods to analyze imprecisely-valued data generated from REs concern 'point' estimation or testing 'two-sided' hypotheses (i.e., null hypotheses consisting of equalities involving population characteristics of the distribution of the REs).

An approach to the *'region' estimation of the fuzzy-valued mean of a one-dimensional RFS* has been stated in González-Rodríguez *et al.* [39]. Given a probability space $(\Omega, \mathcal{A}, P)$, if $X : \Omega \to \mathcal{K}_c(\mathbb{R}^p)$ is an RS, the Aumann mean of X is defined [1] as the set $E^A(X) \in \mathcal{K}_c(\mathbb{R}^p)$ such that

$$E^A(X) = \left\{ \int_\Omega f(\omega)\, dP(\omega) \,:\, f : \Omega \to \mathbb{R}^p,\ f \in L^1(\Omega, \mathcal{A}, P),\ f \in X\ a.s.\,[P] \right\}.$$

If $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R}^p)$ is an RFS, the Aumann-type mean of $\mathcal{X}$ is defined [51] to be given as the fuzzy set $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c^*(\mathbb{R}^p)$ such that for all $\alpha \in [0,1]$

$$\left(\widetilde{E}(\mathcal{X})\right)_\alpha = E^A(\mathcal{X}_\alpha).$$

In case $p = 1$, $\left(\widetilde{E}(\mathcal{X})\right)_\alpha = [E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)]$ for all $\alpha \in [0,1]$.

To estimate $\widetilde{E}(\mathcal{X})$, a 'confidence region' is suggested on the basis of the sample mean of a simple random sample from $\mathcal{X}$, $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ (i.e., $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are independent and identically distributed as $\mathcal{X}$), which is given by

$$\overline{\mathcal{X}_n} = \frac{1}{n} \cdot (\mathcal{X}_1 + \ldots, +\mathcal{X}_n)$$

by looking for a value $\delta > 0$ such that for an arbitrarily fixed confidence level $\tau \in (0,1)$

$$P\left(D_\theta^\varphi\left(\widetilde{E}(\mathcal{X}), \overline{\mathcal{X}_n}\right) \le \delta\right) = \tau,$$

where δ is chosen to be the τ-quantile of the distribution of $D_\theta^\varphi\left(\widetilde{E}(\mathcal{X}), \overline{\mathcal{X}_n}\right)$ which could be approximated by the corresponding boostrap quantile. The confidence region (fuzzy ball) for $\widetilde{E}(\mathcal{X})$ will then be given by

$$\left\{\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R}^p) \,:\, D_\theta^\varphi\left(\widetilde{U}, \overline{\mathcal{X}_n}\right) \le \delta\right\}.$$

In González-Rodríguez *et al.* [39] the approach is illustrated in the one-dimensional case by means of empirical studies based on the simulation ideas from Section 2.

In connection with two-sided hypothesis testing, the k-sample case procedure in [34] has been improved and generalized to *test the equality of Aumann-type means* of k independent RFSs (González-Rodríguez *et al.* [37]). A technique has also been sketched for k dependent RFSs (Montenegro *et al.* [47]), and a deeper study is forthcoming. The generalized one-way ANOVA (ANanlysis Of, VAriance) for independent samples in [39] has been formalized on the basis of the isometrical identification of fuzzy and functional data. Actually, the improved-generalized method is applicable to one-way ANOVA for functional data.

A two-way ANOVA for set-valued data has been stated by considering: two factors, X_{ijk} denoting the k-th response RS ($k \in \{1,\ldots,n_{ij}\}$) under the i-th level of the first factor ($i \in \{1,\ldots,I\}$) and the j-th level of the second factor ($j \in \{1,\ldots,J\}$), and the 'linear' model

$$X_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk},$$

where $\mu, \alpha_i, \beta_j, \delta_{ij} \in \mathcal{K}_c(\mathbb{R}^p)$, and ε_{ijk} are independent RSs. Then, given the null hypothesis of 'no effect of the first factor', that is, $H_0^{(1)} : \alpha_1 = \ldots = \alpha_I$, one can consider the statistic

$$T_n^{(1)} = \frac{\displaystyle\sum_{i=1}^{I}\left(\sum_{j=1}^{J} n_{ij}\right)\left[d_\theta(\overline{X}_{i\cdot\cdot}, \overline{X}_{\cdots})\right]^2}{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}\left[d_\theta(X_{ijk}, \overline{X}_{ij\cdot})\right]^2},$$

with (if $n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$)

$$\overline{X}_{\cdots} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} X_{ijk},\ \overline{X}_{i\cdot\cdot} = \frac{1}{\displaystyle\sum_{j=1}^{J} n_{ij}}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} X_{ijk},\ \overline{X}_{ij\cdot} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} X_{ijk}.$$

A bootstrap approximation of the distribution of the statistic is described in Nakama *et al.* [48] to test $H_0^{(1)}$ at (approximately) an arbitrarily given nominal significance level. Analogous statistics and procedures are established to test the null hypotheses $H_0^{(2)} : \beta_1 = \ldots = \beta_J$ and $H_0^{(1,2)} : \delta_{11} = \ldots = \delta_{IJ}$. A factorial ANOVA for fuzzy set-valued data has been presented in Nakama *et al.* [49].

Asymptotic and bootstrap techniques have been developed to *test two-sided hypotheses about the Fréchet variances of RFSs* by extending classical tests. Given a probability space $(\Omega, \mathcal{A}, P)$ and an RFS $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R}^p)$, the Fréchet variance (more precisely, the θ, φ-Fréchet variance) of $\mathcal{X}$ is defined [37] (extending the definition from the one-dimensional case in Lubiano *et al.* [42] and particularizing to some extent that in [41]) as

$$\sigma^2_{\mathcal{X}} = E\left(\left[D_\theta^\varphi(\mathcal{X}, \widetilde{E}(\mathcal{X}))\right]^2\right).$$

In Ramos-Guajardo *et al.* [52] one-sample asymptotic and bootstrap tests for the Fréchet variance are introduced, and the power function of the asymptotic approach is discussed through local alternatives. The last discussion is extended in Ramos-Guajardo *et al.* [53] to test the homocedasticity of k RFSs. The k-sample case is also analyzed in depth (see Ramos-Guajardo and

Lubiano [54]) by extending Levene's classical procedure, this extension being empirically compared with a Bartlett-type test by considering simulation developments based on the ideas outlined in Section 2.

As an instance of the studies for $k = 2$, in [54], if $\mathcal{X}$ and $\mathcal{Y}$ are two independent RFSs and simple random samples from them, $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ and $(\mathcal{Y}_1, \ldots, \mathcal{Y}_m)$, are available to test the null hypothesis $H_0 : \sigma^2_{\mathcal{X}} = \sigma^2_{\mathcal{Y}}$, one can consider the statistic

$$T_{n,m} = \sqrt{n}\left(\frac{\sum_{i=1}^{n}\left[D^{\varphi}_{\theta}(\mathcal{X}_i, \overline{\mathcal{X}_n})\right]^2}{n-1} - \frac{\sum_{j=1}^{m}\left[D^{\varphi}_{\theta}(\mathcal{Y}_j, \overline{\mathcal{Y}_m})\right]^2}{m-1}\right)$$
$$\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left[D^{\varphi}_{\theta}(\mathcal{X}_i, \overline{\mathcal{X}_n})\right]^2 - \frac{1}{n}\sum_{i=1}^{n}\left[D^{\varphi}_{\theta}(\mathcal{X}_i, \overline{\mathcal{X}_n})\right]^2\right)\right.$$
$$\left.+\frac{1}{m}\sum_{j=1}^{m}\left(\left[D^{\varphi}_{\theta}(\mathcal{Y}_j, \overline{\mathcal{Y}_m})\right]^2 - \frac{1}{m}\sum_{j=1}^{m}\left[D^{\varphi}_{\theta}(\mathcal{Y}_j, \overline{\mathcal{Y}_m})\right]^2\right)\right)^{-1/2}.$$

Under quite general assumptions the asymptotic distribution of the statistic is standard normal.

Another statistical problem for which an approach has been recently proposed in the one-dimensional imprecisely-valued case is that of defining a *robust summary measure of the central tendency of an RE* by extending the notion of median. Because of the lack of a universally acceptable ranking between set (fuzzy set) values, one way to extend this notion is to define the median as a set (fuzzy set) value minimizing a certain mean L^1-distance w.r.t. the RE. In the interval-valued case, two convenient choices have been given by considering the Hausdorff metric (or a generalized version), leading to the interval value(s) $\mathrm{Me}(X) = [\mathrm{Me}(\mathrm{mid}\, X) - \mathrm{Me}(\mathrm{spr}\, X), \mathrm{Me}(\mathrm{mid}\, X) + \mathrm{Me}(\mathrm{spr}\, X)]$ (see Sinova *et al.* [55]), or the 1-norm one (introduced in Sinova *et al.* [57] for RFSs with $p = 1$) leading to the interval values(s) $\mathrm{Me}(X) = [\mathrm{Me}(\inf X), \mathrm{Me}(\sup X)]$. An empirical comparison between the two medians is included in this book [58]. As we have just mentioned, the second approach has been recently extended to RFSs for $p = 1$ leading to a fuzzy number coinciding levelwise with $\left(\widetilde{\mathrm{Me}}(\mathcal{X})\right)_{\alpha} = [\mathrm{Me}(\inf \mathcal{X}_{\alpha}), \mathrm{Me}(\sup \mathcal{X}_{\alpha})]$. However, the first approach cannot be trivially extended because whereas one can trivially state sufficient and necessary conditions for the mid and spr to determine an interval value, one cannot state separately sufficient and necessary conditions for the mid and spr to allow constructing a fuzzy number (see Sinova *et al.* [56]).

4 Summary of Recent Results on Regression Analysis with Imprecisely-valued Data

The SMIRE RG has developed several studies related to *regression analysis involving imprecisely-valued data*. Some of the most recent research developments of the Group have also addressed this problem.

On one hand, studies have been carried out to analyze a flexible simple linear model in which both output (response) and input (explanatory) data are assumed to be interval-valued. This flexible model (referred to as Model M) is formalized by means of the linear relationship

$$Y = a \operatorname{mid} X + b \cdot [-\operatorname{spr} X, \operatorname{spr} X] + c + \varepsilon,$$

where Y is the interval-valued dependent (i.e., the response) RS, X is the interval-valued independent (i.e., the explanatory or regressor) RS, a and b are the real-valued regression coefficients associated with X, c is the real-valued intercept term affecting $\operatorname{mid} Y$, and ε is an interval-valued random error satisfying $E^A(\varepsilon|X) = [-\delta, \delta]$ with $\delta \geq 0$. The definition of the model implies that the errors can be expressed as the Hukuhara distance between the response and the regression function.

In Blanco-Fernández *et al.* [8], arguments supporting the practical interest of model M have been given. Furthermore, Model M has been estimated by using the least squares (LS) approach involving the metric d_θ, which for a given a simple random sample $\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$ from (X, Y) can be stated as follows:

$$\left.\begin{array}{ll} \min\limits_{a,b} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \Big[d_\theta\big(Y_i - Y_i^\circ(a,b), \overline{Y_n} - \overline{Y^\circ(a,b)_n}\big) \Big]^2 \\ \\ \text{subject to} \quad Y_i - Y_i^\circ(a,b) \text{ existing for all } \; i = 1, \ldots, n \end{array}\right\}$$

with $Y_i^\circ(a,b) = a \operatorname{mid} X_i + b \cdot [-\operatorname{spr} X_i, \operatorname{spr} X_i]$, whence $\overline{Y^\circ(a,b)_n} = a\, \overline{(\operatorname{mid} X)_n} + b \cdot [-\overline{(\operatorname{spr} X)_n}, \overline{(\operatorname{spr} X)_n}]$. The constraints in the LS problem are related to the existence of the residuals which, from the definition of the model, should be the Hukuhara distances between the responses and the estimated values from the regression function.

The solutions can be expressed in terms of the models for the real-valued random variables mid and spr of the dependent and independent RSs as follows:

$$\widehat{a} = \frac{\widehat{\sigma_{\operatorname{mid} X, \operatorname{mid} Y}}}{\widehat{\sigma^2_{\operatorname{mid} X}}}, \quad \widehat{b} = \min\left\{ \widehat{s_0}, \max\left\{ 0, \frac{\widehat{\sigma_{\operatorname{spr} X, \operatorname{spr} Y}}}{\widehat{\sigma^2_{\operatorname{spr} X}}} \right\} \right\}$$

$(\widehat{s_0} = \min_i\{\operatorname{spr} Y_i/\operatorname{spr} X_i \ : \ \operatorname{spr} X_i \neq 0\})$ and, hence $\widehat{c} = \overline{\operatorname{mid} Y} - \widehat{a}\,\overline{\operatorname{mid} X}$, $\widehat{\delta} = \overline{\operatorname{spr} Y} - \widehat{b}\,\overline{\operatorname{spr} X}$. These estimators were proven to be strongly consistent.

In Blanco-Fernández *et al.* [6, 9] confidence sets for the parameters of Model M have been stated, and their performance has been empirically investigated. In this book [7] several linear regression techniques for interval-valued data have been revised by comparing their performance under different conditions. In García-Bárzana *et al.* [31], [32] introductory studies on the multiple regression problem for interval-valued data have been developed.

The fuzzy set-valued case has also been examined. The simple linear regression problem between RFSs based on the usual fuzzy arithmetic and the metric D_θ^φ (actually, a more general one introduced by Körner and Näther [41]) has been discussed. The model is not as flexible as model M for interval-valued data, in the sense that the considered model is given by

$$\mathcal{Y} = a \cdot \mathcal{X} + \mathcal{E}$$

where $\mathcal{Y}$, $\mathcal{X}$ and $\mathcal{E}$ are RFSs. In González-Rodríguez *et al.* [35] solutions for the least-squares estimation of the model are given. The advantage of this approach is that it is applicable for general fuzzy data in $\mathcal{F}_c^*(\mathbb{R}^p)$. However, in the particular case of LR fuzzy numbered data, more flexible models has been considered and the associated least squares approaches have been carried out by the SMIRE RG. In fact, Ferraro *et al.* have examined the simple [27] and multiple [29] linear regression model when the response is assumed to be LR fuzzy-valued and the explanatory terms are supposed to be real-valued. In the simple case, the considered model is formalized (in accordance with the notations for LR fuzzy numbers) by means of the (classical) linear relationships

$$\begin{cases} \mathcal{Y}^m = a_m\, X + \ b_m + \varepsilon_m \\ g(\mathcal{Y}^l) = a_l\, X + \ b_l + \varepsilon_l \\ h(\mathcal{Y}^r) = a_r\, X + \ b_r + \varepsilon_r \end{cases}$$

where $\mathcal{Y}$ is the response RFS, X is the real-valued explanatory random variable, g and h are real-valued invertible functions defined on $(0, +\infty)$, the a's and b's are real-valued regression coefficients, and the ε's are real-valued random errors satisfying $E(\varepsilon_m|X) = E(\varepsilon_l|X) = E(\varepsilon_r|X) = 0$.

In [27] and [29] the involved parameters have been estimated by considering the least squares approach using the metric $D_{\mathfrak{l}\mathfrak{r}}$, so that given a simple random sample $\big((X_1, \mathcal{Y}_1), \ldots, (X_n, \mathcal{Y}_n)\big)$ from $(X, \mathcal{Y})$ the solutions are as follows:

$$\widehat{a_m} = \frac{\widehat{\sigma_{X,\mathcal{Y}^m}}}{\widehat{\sigma_X^2}}, \quad \widehat{a_l} = \frac{\widehat{\sigma_{X,g(\mathcal{Y}^l)}}}{\widehat{\sigma_X^2}}, \quad \widehat{a_r} = \frac{\widehat{\sigma_{X,h(\mathcal{Y}^r)}}}{\widehat{\sigma_X^2}},$$

$$\widehat{b_m} = \overline{\mathcal{Y}^m} - \widehat{a_m}\,\overline{X}, \quad \widehat{b_l} = \overline{g(\mathcal{Y}^l)} - \widehat{a_l}\,\overline{X}, \quad \widehat{b_r} = \overline{h(\mathcal{Y}^r)} - \widehat{a_m}\,\overline{X}.$$

Asymptotic properties of the estimators have been analyzed, and confidence regions and tests about the regression parameters and for the linearity have also been established.

In Ferraro *et al.* [28] a determination coefficient, and an associated test for linear independence based on the preceding model, have been introduced. These studies have also been extended to cases with fuzzy-valued explanatory data by Ferraro and Giordani [30]. In Ferraro *et al.* [27] a linearity test between an LR fuzzy-valued response and a real-valued predictor is stated.

A recent approach to robust fuzzy regression analysis has been developed by D'Urso *et al.* [26]. The study considers an extension of a previous model by Coppi *et al.* [17] consisting of the (classical) linear relationships

$$\begin{cases} \mathcal{Y}^m = a_1\, f_1(X_1) + \ldots + a_k\, f_k(X_k) + \varepsilon \\ \inf \mathcal{Y}_1 = b(a_1\, f_1(X_1) + \ldots + a_k\, f_k(X_k)) + c + \varepsilon_{\inf} \\ \sup \mathcal{Y}_1 = d(a_1\, f_1(X_1) + \ldots + a_k\, f_k(X_k)) + e + \varepsilon_{\sup} \end{cases}$$

where $\mathcal{Y}$ is the response RFS, $X_1, \ldots, X_k$ are the real-valued explanatory random variables, f_j is a real-valued function representing the regression 'profile' of the observation in terms of a suitably chosen function of X_j, $(a_1, \ldots, a_k)$ is the vector of coefficients of the regression model for $\mathcal{Y}^m$, b, c, d and e are real-valued regression coefficients, and the ε's are the residuals of the models. It should be noted that the linear models for the inf and sup of the 1-level of the response RFS are constructed on the basis of the linear model for $\mathcal{Y}^m$, allowing for possible (classical) linear relationships between the magnitude of the estimated left and right deviations, $\mathcal{Y}^l$ and $\mathcal{Y}^r$, and that of the estimated $\mathcal{Y}^m$. To estimate the model a robust procedure (avoiding the concerns related to possible outliers) has been suggested by replacing the least squares criterion by either the least median squares (i.e., the median of the squared residuals) or the weighted least squares (assigning low weights to data identified as outliers) on the basis of the D_{lr} metric. After explaining how to estimate the parameters in the model with each of the two approaches, a determination coefficient is introduced, and simulations show empirically that in the presence of outliers the robust method outperforms the least squares approach.

In D'Urso *et al.* [25] a method has been suggested to cluster data into rather homogeneous groups to subsequently fit separate least squares regression analysis (i.e., a different regression model) within each cluster.

5 Summary of Recent Results on Clustering/Classification of Imprecisely-valued Data

Another problem which the SMIRE RG has tried to tackle during the last years is that of *clustering imprecisely-valued data* into groups or providing

criteria to classify a given imprecisely-valued datum into one category of a set of pre-established ones.

For the case of interval-valued data D'Urso and De Giovanni [21] have introduced a method to cluster units which are identified with 'vectors' of interval-valued data, by using an unsupervised neural network approach. For this purpose, a distance between units introduced by D'Urso and Giordani [22] is exploited through the so-called midpoint radius self-organizing maps. The effectiveness of the new method is discussed and some applications are shown.

In case of fuzzy-valued data a clustering method has been suggested by González-Rodríguez *et al.* [36]. The goal has been to group k independent RFSs by focusing on their fuzzy means. The procedure is iterative and it has been based on the multi-sample (ANOVA) bootstrap tests in [34] and [37]. For an arbitrarily fixed significance level the method determines groups as follows: for any two different clusters there exist elements in them for which the fuzzy means are significantly different; and for all RFSs in the same cluster their fuzzy means could be considered to coincide (i.e., there is not enough sampling evidence to conclude rejection of the null hypothesis of equality of the means) at the given significance level. An objective stopping criterion leading to statistically equal groups different from each other has also been presented, and simulations have been carried out (following ideas at the end of Section 2) to show the performance of the suggested procedure.

In Coppi *et al.* [16] the well-known fuzzy k-means clustering model has been, on one hand, adapted to deal with LR fuzzy numbered data and, on the other hand, relaxed by removing an orthogonality assumption. The models and methods (called, respectively, fuzzy k-means and possibilistic k-means for fuzzy data) are based on a new metric strongly related to $D_{\mathfrak{lr}}$. They allow us to detect k homogeneous clusters on the basis of n objects described by means of several RFSs (with $p = 1$) with LR fuzzy numbered values. By means of empirical and practical developments, some first conclusions have been drawn in connection with the advantages of the use of the possibilistic approach.

In connection with the classification of fuzzy-valued data into one of a set of categories, a discriminant approach based on nonparametric kernel density estimation is introduced in Colubi *et al.* [13, 14]. Since the procedure is shown not to be optimal in general and to require large sample sizes, a simpler approach eluding the density estimation is proposed in Colubi *et al.* [13]. Assume that individuals to be classified may belong to one of k different categories $G_1, \ldots, G_k$, and as learning sample (a supervised classification approach is followed) we have a group of n independent individuals and the corresponding fuzzy data corresponding to an RFS $\mathcal{X}$. The goal is to find a rule allowing us to classify a new individual in one of the k groups on the basis of the associated fuzzy datum, say $\widetilde{x} \in \mathcal{F}_c^*(\mathbb{R}^p)$.

Formally, let $(\mathcal{X}, G) : \Omega \to \mathcal{F}_c^*(\mathbb{R}^p) \times \{1, \ldots, k\}$ be a random element in such a way that $\mathcal{X}(\omega)$ is a fuzzy datum and $G(\omega)$ is the group individual

$\omega \in \Omega$ belongs to (i.e., $G(\omega) \in \{G_1, \ldots, G_k\}$). Assume that we have n independent copies of $(\mathcal{X}, G)$ as training sample, that is, we have a simple random sample of size n, $\{\mathcal{X}_{i,g}\}_{i,g}$ ($i \in \{1, \ldots, n_g\}$, $g \in \{1, \ldots, k\}$, $n_1 + \ldots + n_k = n$). The so-called ball-based classification criterion for fuzzy data consists of computing first the distances between $\widetilde{x}$ and the training fuzzy data by using the metric D_θ^φ, that is,

$$d_{i,g} = D_\theta^\varphi(\widetilde{x}, \mathcal{X}_{i,g}).$$

Later, for a chosen value $\delta > 0$ and for each group one computes $n_{\delta,g} = \sum_{i=1}^{n_g} \mathbf{1}_{[0,\delta]}(d_{i,g})$. Then, one should estimate the probabilities of belonging to each group by means of the numbers

$$\widehat{P}\Big(G = G_g \,\Big|\, \mathcal{X} \in \Big\{\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R}^p) \,:\, D_\theta^\varphi(\widetilde{U}, \widetilde{x}) \leq \delta\Big\}\Big) = \frac{n_{\delta,g}}{\sum_{h=1}^{k} n_{\delta,h}}.$$

Finally, $\widetilde{x}$ will be assigned to the group associated with the highest estimated probability. Some comparative developments have been carried out in [13], in which the new methods have been compared with linear discriminant analysis and random K-fold cross validation.

Although they do not correspond to clustering or classification of imprecisely-valued data, but to crisp ones, it can be remarked in this context that several fuzzy techniques (i.e., leading to fuzzy clusters or categories) have been recently introduced by SMIRE RG members. In this respect one can mention, as related to the COST Action IC0702, the contributions by Coppi *et al.* [15], D'Urso and Maharaj [23, 24], Maharaj and D'Urso [44] and Maharaj *et al.* [45].

6 SAFD (Statistical Analysis of Fuzzy Data) R Package

Recently, an R package has been designed (see, for instance, Lubiano and Trutschnig [43] and Trutschnig and Lubiano in this book [61]) to provide some basic functions for doing statistics with one-dimensional fuzzy data. In particular, the package contains functions for the basic operations on the class of fuzzy numbers (sum, scalar product, mean, Hukuhara difference, Aumann-type mean, 1-norm median, etc.) as well as for calculating some distances, sample Fréchet variance, sample covariance, sample correlation, and so on. Moreover a function to simulate fuzzy random variables, bootstrap tests for the equality of means, and a function to do linear regression given trapezoidal fuzzy data is included. The package is being almost permanently updated by incorporating most of the new developed methods, and ease substantially the computing process in the applications of these methods and in the empirical studies.

7 Concluding Remarks

A summary of recent statistical developments for imprecise data from the ontic point of view has been done. Imprecise data are here formalized through intervals and fuzzy sets. The common points of the different approaches are that they are based on the (fuzzy) set-arithmetic and a scalar distances, since the aim is to handle each data as an entity. Many of the statistics are connected with estimation and testing procedures about parameters and can be applied by using a freely available R package. Further details and discussions can be found in the referred works.

From a technical point of view, it should be underlined that the procedures in this paper concerning fuzzy-valued data have been presented (for the sake of simplicity) for the space $\mathcal{F}_c^*(\mathbb{R}^p)$, although they are valid for a wider space allowing us to identify each fuzzy value with a function within a Hilbert space (as pointed out in [5] and [37]).

SMIRE researchers have also been involved in research directions that do not fit the scope of this review which, however, are strongly related to the COST Action IC0702 too. The webpage of the RG references these works (see `http://bellman.ciencias.uniovi.es/SMIRE/Publications.html`).

Acknowledgements This research has been partially supported by/benefited from several Short-Term Scientific Missions (Blanco-Fernández, Colubi, De la Rosa de Sáa, Ferraro, García-Bárzana, González-Rodríguez, Ramos-Guajardo and Sinova) and Working Group Meetings associated with the COST Action IC0702, as well as by the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02. Their financial support is gratefully acknowledged.

References

1. Aumann R (1965) Integrals of set-valued functions. *J Math Anal Appl* 12:1–12
2. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. *Mathware Soft Comp* 2:71–84
3. Borgelt C, González-Rodríguez G, Trutschnig W, Lubiano MA, Gil MA, Grzegorzewski P, Hryniewicz O, eds. (2010) *Combining Soft Computing and Statistical Methods in Data Analysis*. Springer-Verlag, Heidelberg
4. Borgelt C, Gil MA, Sousa JMC, Verleysen M, eds. (2012) *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, to appear. Springer-Verlag, Heidelberg
5. Blanco-Fernández A, Casals MR, Colubi A, Corral N, García-Bárzana M, Gil MA, González-Rodríguez G, López MT, Lubiano MA, Montenegro M, Ramos-Guajardo AB, de la Rosa de Sáa S, Sinova B (2012) Random fuzzy sets: a mathematical tool to develop statistical fuzzy data analysis. *Iran J Fuzzy Syst.*, in press
6. Blanco-Fernández A, Colubi A, González-Rodríguez G (2012) Confidence sets in a linear regression model for interval data. *J Statist Planning Infer*, 142(6):1320–1329
7. Blanco-Fernández A, Colubi A, González-Rodríguez G (2012) Linear regression analysis for interval data based on set arithmetic: a review. In [4], 19–32

8. Blanco-Fernández A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. *Comp Stat Data Anal* 55:2568–2578
9. Blanco-Fernández A, Corral N, González-Rodríguez G, Palacio A (2010) On some confidence regions to estimate a linear regression model for interval data. In [3], 33–40
10. Colubi A, Domínguez-Menchero JS, López-Díaz M, Ralescu DA (2001) On the formalization of fuzzy random variables. *Inform Sci* 133:3–6
11. Colubi A, Domínguez-Menchero JS, López-Díaz M, Ralescu DA (2002) A $D_E[0,1]$ representation of random upper semicontinuous functions. *Proc Am Math Soc* 130:3237–3242
12. Colubi A, Fernández C, Gil MA (2002) Simulation of random fuzzy variables: an empirical approach to statistical/probabilistic studies with fuzzy experimental data. *IEEE Trans Fuzzy Syst* 10:384–390
13. Colubi A, González-Rodríguez G, Gil MA, Trutschnig W (2011) Nonparametric criteria for supervised classification of fuzzy data. *Int J Approx Reas* 52:1272–1282
14. Colubi A, González-Rodríguez G, Trutschnig W (2009) Discriminant analysis for fuzzy random variables based on nonparametric regression. *Abst IFSA/EUSFLAT Conf*
15. Coppi R, D'Urso P, Giordani P (2010) A fuzzy clustering model for multivariate spatial time series. *J Classif* 27:54–88
16. Coppi R, D'Urso P, Giordani P (2012) Fuzzy and possibilistic clustering for fuzzy data. *Comp Stat Data Anal*, 56(4):915–927
17. Coppi R, D'Urso P, Giordani P, Santoro A (2006) Least squares estimation of a linear regression model with *LR* fuzzy response. *Comp Stat Data Anal* 51:267–286
18. Diamond P, Kloeden P (1999) Metric spaces of fuzzy sets. *Fuzzy Sets Syst* 100:63–71
19. Dubois D, Prade H (1978) Operations on fuzzy numbers. *Int. J. of Systems Science* 9(6):613–626
20. Dubois D, Prade H (2012) Gradualness, uncertainty and bipolarity: making sense of fuzzy sets. *Fuzzy Sets Syst* 192:3–24
21. D'Urso P, De Giovanni L (2011) Midpoint radius self-organizing maps for interval-valued data with telecommunications application. *Appl Soft Comp* 11:3877–3886
22. D'Urso P, Giordani P (2004) A least squares approach to principal component analysis for interval valued data. *Chem Intel Lab Syst* 70:179–192
23. D'Urso P, Maharaj EA (2009) Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets Syst* 160:3565–3589
24. D'Urso P, Maharaj EA (2012) Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems* 193:33–61
25. D'Urso P, Massari R, Santoro A (2010) A class of fuzzy clusterwise regression models. *Inform Sci* 180:4737–4762
26. D'Urso P, Massari R, Santoro A (2011) Robust fuzzy regression analysis. *Inform Sci* 181:4154–4174
27. Ferraro MB, Colubi A, Giordani P (2010) A linearity test for a simple regression model with *LR* fuzzy response. In [3], 251–258
28. Ferraro MB, Colubi A, González-Rodríguez G, Coppi R (2011) A determination coefficient for a linear regression model with imprecise response. *Environmetrics* 22:516–529
29. Ferraro MB, Coppi R, González-Rodríguez G, Colubi A (2010) A linear regression model for imprecise response. *Int J Approx Reas* 51:759–770
30. Ferraro MB, Giordani P (2012) A multiple linear regression model for imprecise information. *Metrika*, in press (doi:10.1007/s00184-011-0367-3)
31. García-Bárzana M, Colubi A, Kontoghiorghes E (2010) Least-squares estimation of a multiple regression model for interval data. *Abst 3rd Workshop ERCIM'10.*
32. García-Bárzana M, Colubi A, Kontoghiorghes E (2011) A flexible multiple linear regression model for interval data. *Abst 4th Workshop ERCIM'11.*

33. Gil MA, Lubiano MA, Montenegro M, López-García MT (2002) Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika* 56:97–111
34. Gil MA, Montenegro M, González-Rodríguez G, Colubi A, Casals MR (2006) Bootstrap approach to the multi-sample test of means with imprecise data. *Comp Stat Data Anal* 51:148–162
35. González-Rodríguez G, Blanco A, Colubi A, Lubiano MA (2009) Estimation of a simple linear regression model for fuzzy random variables. *Fuzzy Sets Syst* 160(3):357–370
36. González-Rodríguez G, Colubi A, D'Urso P, Montenegro M (2009) Multi-sample test-based clustering for fuzzy random variables. *Int J Approx Reas* 50:721–731
37. González-Rodríguez G, Colubi A, Gil MA (2012) Fuzzy data treated as functional data. A one-way ANOVA test approach. *Comp Stat Data Anal.* 56(4):943–955
38. González-Rodríguez G, Colubi A, Trutschnig W (2009) Simulation of fuzzy random variables. *Inform Sci* 179:642–653
39. González-Rodríguez G, Trutschnig W, Colubi A (2009) Confidence regions for the mean of a fuzzy random variable. *Abst IFSA/EUSFLAT Conf*
40. Klement EP, Puri ML, Ralescu DA (1986) Limit theorems for fuzzy random variables. *Proc R Soc Lond A* 407:171–182
41. Körner R, Näther W (2002) On the variance of random fuzzy variables. In: Bertoluzza C, Gil MA, Ralescu DA (eds.) *Statistical Modeling, Analysis and Management of Fuzzy Data*, 22–39. Physica-Verlag, Heidelberg
42. Lubiano MA, Gil MA, López-Díaz M, López MT (2000) The $\overrightarrow{\lambda}$-mean squared dispersion associated with a fuzzy random variable. *Fuzzy Sets Syst* 111:307–317
43. Lubiano MA, Trutschnig W (2010) ANOVA for fuzzy random variables using the R-package SAFD. In [3], 449–456
44. Maharaj EA, D'Urso P (2011) Fuzzy clustering of time series in the frequency domain. *Inform Sci* 181:1187–1211
45. Maharaj EA, D'Urso P, Galagedera D (2010) Wavelets-based fuzzy clustering of time series. *J Classif* 27:231–275
46. Molchanov I (2005) *Theory of Random Sets*. Springer-Verlag, London
47. Montenegro M, López MT, Lubiano MA, González-Rodríguez G (2009) A dependent multi-sample test for fuzzy means. *Abst 2nd Workshop ERCIM'09*.
48. Nakama T, Colubi A, Lubiano MA (2010) Two-way analysis of variance for interval-valued data. In [3], 475–482
49. Nakama T, Colubi A, Lubiano MA (2010) Factorial analysis of variance for fuzzy data. *Abst 3rd Workshop ERCIM'10*.
50. Puri ML, Ralescu DA (1983) Strong Law of Large Numbers for Banach space valued random sets. *Ann Probab* 11:222–224
51. Puri ML, Ralescu DA (1986) Fuzzy random variables. *J Math Anal Appl* 114:409–422
52. Ramos-Guajardo AB, Colubi A, González-Rodríguez G, Gil MA (2010) One sample tests for a generalized Fréchet variance of a fuzzy random variable. *Metrika* 71:185–202
53. Ramos-Guajardo AB, González-Rodríguez G, Montenegro M, López MT (2010) Power analysis of the homoscedasticity test for random fuzzy sets. In [3], 537–544
54. Ramos-Guajardo AB, Lubiano MA (2012) K-sample tests for equality of variances of random fuzzy sets. *Comp Stat Data Anal.* 56(4):956–966
55. Sinova B, Casals MR, Colubi A, Gil MA (2010) The median of a random interval. In [3], 575–583
56. Sinova B, de la Rosa de Sáa S, Gil MA (2012) A generalized L^1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data. Submitted.
57. Sinova B, Gil MA, Colubi A, Van Aelst S (2012) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets Syst.*, in press (doi:10.1016/j.fss.2011.11.004)

58. Sinova B, Van Aelst S (2012) Comparing the medians of a random interval defined by means of two different L^1 metrics. In [4], to appear
59. Trutschnig W (2010) Characterization of the sendograph-convergence of fuzzy sets by means of their L_p- and levelwise convergence. *Fuzzy Sets Syst* 161:1064–1077
60. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Inform Sci* 179(23):3964–3972
61. Trutschnig W, Lubiano MA, Lastra J (2012) SAFD — An R package for Statistical Analysis of Fuzzy Data. In [4], 107–118
62. Vitale RA (1985) Metrics for compact, convex sets. *J Approx Theo* 45:280–287
63. Yang M-S, Ko C-H (1996) On a class of fuzzy c-numbers clustering procedures for fuzzy data. *Fuzzy Sets Syst* 84:49–60
64. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Part 1. *Inform. Sci.* 8:199–249; Part 2. *Inform. Sci.* 8:301–353; Part 3. *Inform. Sci.* 9:43–80

Linear Regression Analysis for Interval-valued Data Based on Set Arithmetic: A Review

Angela Blanco-Fernández[1], Ana Colubi[1], and Gil González-Rodríguez[1]

Abstract When working with real-valued data regression analysis allows to model and forecast the values of a random variable in terms of the values of either another one or several other random variables defined on the same probability space. When data are not real-valued, regression techniques should be extended and adapted to model simply relationships in an effective way. Different kinds of imprecision may appear in experimental data: uncertainty in the quantification of the data, subjective measurements, perceptions, to name but a few. Compact intervals can be effectively used to represent these imprecise data. Set- and fuzzy-valued elements are also employed for representing different kinds of imprecise data. In this paper several linear regression estimation techniques for interval-valued data are revised. Both the practical applicability and the empirical behaviour of the estimation methods is studied by comparing the performance of the techniques under different population conditions.

1 Introduction

Over the last years, the consideration of different sources of imprecision in generating and modeling experimental data have implied the development of advanced statistical and soft computing techniques capable of dealing with these kinds of data. Interval-valued data are considered as a first attempt to extend the treatment of characteristics for which values are real numbers to more flexible scenarios. For instance, when the characteristic has a great uncertainty in the quantification of its values, it may be suitable to formalize these values as intervals instead of real numbers, in order to take into account this uncertainty in the statistical process (see, for instance, [11]). Grouped or

[1] Department of Statistics and Operational Research, University of Oviedo, Spain, blancoangela@uniovi.es · colubi@uniovi.es · gil@uniovi.es

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 19–31.
springerlink.com

censored data are also usually represented by means of intervals (see [10]). Furthermore, intervals can also correspond to the values of attributes which are essentially interval-valued; this is the case when considering measurements of the fluctuation of a certain characteristic during a period of time, ranges of a certain variable, and so on; see, for instance, [4], [5].

Regression analysis problems with interval-valued data have been deeply investigated. In particular, several linear models for relating two or more experimental measurements with interval values have been proposed in the literature over the last years. Some works develop interval techniques on the basis of the so-called *Symbolic Data Analysis* (see [2]). Symbolic interval variables are mainly used for modelling aggregated data or interval descriptions of technical specifications. The Symbolic regression problems are usually solved separately for real-valued variables associated with the intervals, such as the lower and upper limits, or the midpoints and ranges (see [12] and references therein). The resolution of the regression estimation by means of classical techniques does not forbid the possibility of anomalous results such as forecast intervals whose lower bounds are larger than their upper ones. In [12] non-negativity conditions for the regression parameters that forbid such anomalies are included in the estimation process. However, in this case the estimation is solved by means of numerical optimization methods, and no analytical expressions for the regression estimators are obtained. Since no probabilistic assumptions for the regression model are considered, and numerical techniques are employed, the estimation process is just a fitting problem, so the study of statistical properties of the estimators and inferential studies about the model make no sense in this setting. An alternative approach for interval regression is based on the formalization of a linear relationship between interval-valued variables, as a natural generalization of the classical linear models between real-valued variables. Some works in this domain are [3], [4], [5], [6], [7], among others. In this case, regression models between two or more interval-valued variables are formalized in terms of the interval arithmetic. This is the approach we focus on hereafter. The extension of these interval regression models to other kinds of imprecise-valued variables, when working with set- or fuzzy-valued data, can be formalized in a direct way by considering set- or fuzzy- arithmetic (see [8]).

In this paper, the main properties of several interval linear models based on the set arithmetic which have been introduced recently are analyzed; differences among them are also examined. The rest of the paper is organized as follows: in Section 2 some preliminary concepts about the interval framework are presented. In Section 3 different linear models between interval-valued data based on set arithmetic are revised. Their theoretical features and the estimation of their parameters are shown. Results comparing the empirical performance and the practical applicability of the considered models by means of some simulation studies and several practical data sets are gathered in Section 4. Finally, Section 5 states some conclusions and future research.

2 Preliminaries

The statistical treatment of interval-valued experimental data is developed by considering them as elements belonging to the space $\mathcal{K}_c(\mathbb{R}) = \{[a,b] : a,b \in \mathbb{R},\ a \leq b\}$. Each compact interval $A \in \mathcal{K}_c(\mathbb{R})$ can be expressed by means of its (inf, sup)-representation, i.e. $A = [\inf A, \sup A]$, with $\inf A \leq \sup A$. Alternatively, the notation $[\operatorname{mid} A \pm \operatorname{spr} A]$ with $\operatorname{spr} A \geq 0$, where $\operatorname{mid} A = (\sup A + \inf A)/2$ is the *midpoint* of the interval, and $\operatorname{spr} A = (\sup A - \inf A)/2$ denotes the *spread* or *radius* of A, can be considered. The interval A and $[\operatorname{mid} A \pm \operatorname{spr} A]$ are obviously equivalent. Statistical developments with interval-valued data are generally based on the *(mid,spr)*-parametrization, since the non-negativity condition for the *spr* component is usually easier to handle than the order condition for the *inf* and *sup* components of the *(inf, sup)*-characterization.

In order to manage intervals a natural arithmetic is defined on $\mathcal{K}_c(\mathbb{R})$ by means of the Minkowski addition $A + B = \{a + b : a \in A, b \in B\}$ and the product by scalars $\lambda A = \{\lambda a : a \in A\}$, for any $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. The space $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear but semilinear due to the lack of symmetric element with respect to the addition; the operation $A + (-1)B$ does not satisfy, in general, the property that $B + (A + (-1)B) = A$. For example, $[0,2] + (-1)[1,5] = [-5,1]$ and $[1,5] + [-5,1] = [-4,6] \neq [0,2]$. Moreover, in many cases there exists no interval C such that $B + C = A$ (and so $A - B = C$); for instance, for $A = [1,2]$ and $B = [0,4]$, the unique way to get $B + C = A$ is being $C = [1,-2] \notin \mathcal{K}_c(\mathbb{R})$. Thus, the difference $A - B$ is not an inner operation in the space $\mathcal{K}_c(\mathbb{R})$. When an element $C \in \mathcal{K}_c(\mathbb{R})$ such that $B + C = A$ exists, then it is the so-called *Hukuhara difference* between intervals A and B. The existence of C is assured if $\operatorname{spr} B \leq \operatorname{spr} A$. In that case, it is denoted $C = A -_H B$.

In order to measure distances between two intervals, there are several metrics defined on $\mathcal{K}_c(\mathbb{R})$ (see, for instance, [1], [13]) . For regression problems, and in particular for least squares methods, in which distances are employed for error measurements, an L_2-type metric has been exhaustively used and shown to be suitable. The so-called d_θ-*distance* defined in [13] as

$$d_\theta(A,B) = \sqrt{(\operatorname{mid} A - \operatorname{mid} B)^2 + \theta(\operatorname{spr} A - \operatorname{spr} B)^2}\ , \tag{1}$$

for an arbitrarily chosen $\theta > 0$, generalizes all the metrics employed in the estimation process of the linear models examined in Section 3.

2.1 Random Intervals

In the probabilistic setting, the modeling of random elements with experimental interval values is based on the notion of a compact random interval (or

interval-valued random set). The concept of a random interval is formalized as follows:

Definition 1. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A mapping $X: \Omega \to \mathcal{K}_c(\mathbb{R})$ is said to be a random interval (or, more precisely, a compact random interval) if it is $\mathcal{B}_{d_\theta}|\mathcal{A}$-measurable, $\mathcal{B}_{d_\theta}$ denoting the σ-field generated by the topology induced by the metric d_θ on $\mathcal{K}_c(\mathbb{R})$.

Equivalently, the definition of random interval for the mapping X can be formalized through the associated real-valued mappings (see [14]):

(i) $\inf X, \sup X : \Omega \to \mathbb{R}$, being random variables and $\inf X \leq \sup X$ a.s.-[P][1]
(ii) $\operatorname{mid} X, \operatorname{spr} X : \Omega \to \mathbb{R}$, being random variables and $\operatorname{spr} X \geq 0$ a.s.-[P]

Remark 1. As mentioned already, the non-negativity condition for the second component of the real random vector $(\operatorname{mid} X, \operatorname{spr} X)$ in (ii) is usually easier to handle in statistical processing than the order condition in (i), so the second characterization for X will be used.

Analogously to the classical scenario, some summary measures and parameters can be defined for a random interval X. The most commonly used definition for the expected value of X is based on the well-known Aumann expectation of imprecise random elements. In case of X being a random interval, it can be expressed as

$$E(X) = [E(\operatorname{mid} X) \pm E(\operatorname{spr} X)], \tag{2}$$

whenever the involved expected values exist, i.e. $\operatorname{mid} X, \operatorname{spr} X \in L^1$. This concept satisfies the usual properties of linearity, and it is the *Fréchet expectation* w.r.t. d_θ (see [9]). This allows to define the variance of X as the usual *Fréchet variance* associated with the Aumann expectation in the metric space $(\mathcal{K}_c(\mathbb{R}), d_\theta)$, i.e.,

$$\sigma_X^2 = d_\theta^2(X, E(X)) \ , \tag{3}$$

whenever $E(|X|^2) < \infty$ (where $|X| = \sup\{|x| : x \in X\}$). This concept verifies also the usual properties for a variance of a random variable. Moreover, it can be expressed in terms of classical variances as $\sigma_X^2 = \sigma^2_{\operatorname{mid} X} + \theta\sigma^2_{\operatorname{spr} X}$. The semilinearity of the space $\mathcal{K}_c(\mathbb{R})$ entails some difficulties to extend the classical concept of covariance for two random intervals X and Y. Thus, $\sigma_{X,Y}$ is often defined as the corresponding d_θ-covariance in $\mathbb{R}^2$ through the (mid, spr)-parametrization of the intervals, leading to the expression

$$\sigma_{X,Y} = \sigma_{\operatorname{mid} X, \operatorname{mid} Y} + \theta\sigma_{\operatorname{spr} X, \operatorname{spr} Y} \ , \tag{4}$$

whenever the corresponding moments for the involved random variables *mid* and *spr* exist. Although this concept preserves many of the properties of

[1] The notation *a.s.-[P]* corresponds to *almost sure with respect to probability P.*

classical covariance, the ones related to linear transformations of the random intervals may fail; for instance, $\sigma_{aX,Y} \neq a\sigma_{X,Y}$ in general.

The corresponding sample moments for X and Y can be defined in the usual way when a simple random sample $\{X_i, Y_i\}_{i=1}^n$ is obtained from (X, Y). Namely, $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$, $\widehat{\sigma}_X^2 = \frac{1}{n}(\sum_{i=1}^n d_\theta^2(X_i, \overline{X})) = \widehat{\sigma}_{\text{mid}\,X}^2 + \theta\widehat{\sigma}_{\text{spr}\,X}^2$ (analogously for Y) and $\widehat{\sigma}_{X,Y} = \widehat{\sigma}_{\text{mid}\,X,\text{mid}\,Y} + \theta\widehat{\sigma}_{\text{spr}\,X,\text{spr}\,Y}$.

3 Simple Interval Linear Models Based on Set Arithmetic

The formalization of a model for linearly relating two random intervals has been introduced in different ways in the recent literature. In this section some of the main models in this context are examined, showing the advantages and drawbacks of each of them.

One of the first studies on the estimation of a simple linear model for interval-valued data has been developed by Diamond in 1990 (see [4]). This study consists of a linear fitting problem for a given sample interval data set, without probabilistic assumptions for the data. Moreover, the solution for the problem is assured only if the interval data set satisfies some particular conditions which become very restrictive in most of the practical situations (see [5] for details). For this reason, this approach will not be included in the comparative study in Section 4. However, it is worthy to present it as one of the first attempts to address the linear regression fitting for intervals and being an inspiration for the following studies in other works.

3.1 Simple Basic Linear Model without Constraints: Gil et al. (2002)

The linear model introduced in [5] was formalized in a probabilistic context. The aim is to determine the best approximation of a random interval Y by an affine function of another random interval X. A least squares method based on the well-known *Bertoluzza metric* introduced in [1] (which is equivalent to d_θ with $\theta \leq 1$) is developed.

The regression problem is formalized as follows: let $X, Y : \Omega \to \mathcal{K}_c(\mathbb{R})$ be two random intervals associated with $(\Omega, \mathcal{A}, P)$, and let $\{X_i, Y_i\}_{i=1}^n$ be a random sample obtained from (X, Y). The objective is to minimize the function $\phi : \mathbb{R} \times \mathcal{K}_c(\mathbb{R}) \to [0, \infty)$ given by

$$\phi(b, C) = \frac{1}{n}\sum_{i=1}^{n} d_W^2(Y_i, bX_i + C), \tag{5}$$

where d_W denotes the Bertoluzza metric for intervals.

The optimal solutions for this minimization problem are obtained in [5] by means of an algorithm, leading to different expressions depending on the combination of different possible sample situations. Thus, unified analytical expressions for the optimal values of b and C in (5) are not given. Moreover, although the regression problem is established in a probabilistic setting, in the resolution of (5) no information about the probabilistic model from which the sample data are generated is considered. It can be shown that, contrary to what happens in the real-valued case, the fitting model addressed in [5] can be unsuitable for estimating the linear theoretical relationship between X and Y (see [7]). This is due to the special features of the interval arithmetic. A simple example to illustrate this fact is described below.

Example 1. Let X, ε be random intervals characterized through the real variables $\operatorname{mid} X, \operatorname{mid} \varepsilon \sim N(0,1)$, independent of each other, and $\operatorname{spr} X, \operatorname{spr} \varepsilon$ independent χ^2_1 variables. If Y is defined as $Y = X + \varepsilon$, then $E(Y|x) = x + [-1,1]$, for all $x \in \operatorname{Im}(X) \subset \mathcal{K}_c(\mathbb{R})$. Table 1 contains simulated data of $n = 3$ individuals from this situation. The solution to the fitting problem (5) proposed in [5] from this sample data is $b^* = 2.2644$ and $C^* = [-0.3282, 0.4041]$. Since $\operatorname{spr}(b^* X_1) = b^* \operatorname{spr} X_1 = 0.6483 > \operatorname{spr} Y_1 = 0.3756$, the Hukuhara difference $Y_1 -_H (b^* \operatorname{spr} X_1)$ is not defined, i.e. there exists no $\varepsilon_1^* \in \mathcal{K}_c(\mathbb{R})$ such that $Y_1 = b^* X_1 + \varepsilon_1^*$. Thus, it is not possible to reproduce the theoretical linear model from which the data came through the optimal affine function.

Table 1 Sample data set from the linear model $E(Y|x) = x + [-1, 1]$

$\operatorname{mid} X_i$	$\operatorname{spr} X_i$	$\operatorname{mid} Y_i$	$\operatorname{spr} Y_i$
0.6561	0.2863	0.7799	0.3756
-0.0334	0.0653	1.0602	0.8987
-0.2719	0.5166	-2.0318	1.9041

3.2 Simple Basic Linear Model with Constraints: González-Rodríguez et al. (2007)

To overcome the difficulties of the optimal affine function obtained in [5], in [6] a restricted least squares method to estimate the linear relation between two random intervals has been proposed. Afterwards, the estimation problem associated with this regression model has been solved in [7].

Let $X, Y : \Omega \to \mathcal{K}_c(\mathbb{R})$ be two random intervals related by means of a **simple basic linear model** as

$$Y = aX + \varepsilon \ , \tag{6}$$

where ε is an interval random error such that $E(\varepsilon|X) = B \in \mathcal{K}_c(\mathbb{R})$. Given a simple random sample $\{X_i, Y_i\}_{i=1}^n$ from (X, Y), any pair of intervals (X_i, Y_i) fulfills the linear model (6), so $Y_i = aX_i + \varepsilon_i$ for a certain ε_i. Thus, $\varepsilon_i = Y_i -_H aX_i$ for all $i = 1, \ldots, n$.

Remark 2. The interval linear model (6) tries to mimic the classical simple linear model between two real-valued random variables, with some particularities in order to keep the coherency with the interval arithmetic. The (interval-valued) independent term B is included as the expectation of the error in order to allow the error to be an interval-valued random set. It is straightforward to show that if the alternative interval linear model $Y = aX+B+\varepsilon$ such that $E(\varepsilon|X) = 0$ is considered, then ε degenerates to a real-valued random variable. Thus, in order to consider interval-valued errors, the independent term is included in the formalization of the possible errors.

Analogously to the process in [5], the estimation of the regression parameters (a, B) is addressed by means of a least squares criterion with respect to Bertoluzza metric, i.e. minimizing the function $\phi(b, C)$ in (5) for $b \in \mathbb{R}$ and $C \in \mathcal{K}_c(\mathbb{R})$. Moreover, the optimum values are expected to satisfy the linear model (6) at least for the sample intervals. Thus, the minimization problem associated with the estimation of (6) is formalized as follows:

$$\left.\begin{array}{l} \min\limits_{b\in\mathbb{R}, C\in\mathcal{K}_c(\mathbb{R})} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} d_W^2(Y_i, bX_i + C) \\ \text{subject to} \\ Y_i -_H bX_i \ \text{ exists, for all } i = 1, \ldots, n \end{array}\right\} \tag{7}$$

In [7] this contrained minimization problem has been solved, and the following analytical expressions for the estimates of the regression parameters (a, B) have been obtained:

$$a^* = \begin{cases} 0 & \text{if } \widehat{\sigma}_{X,Y} \leq 0 \text{ and } \widehat{\sigma}_{-X,Y} \leq 0 \\ \min\left\{\widehat{s}_0, \dfrac{\widehat{\sigma}_{X,Y}}{\widehat{\sigma}_X^2}\right\} & \text{if } \widehat{\sigma}_{X,Y} \geq 0 \text{ and } \widehat{\sigma}_{-X,Y} \leq \widehat{\sigma}_{X,Y} \\ -\min\left\{\widehat{s}_0, \dfrac{\widehat{\sigma}_{-X,Y}}{\widehat{\sigma}_X^2}\right\} & \text{if } \widehat{\sigma}_{-X,Y} \geq 0 \text{ and } \widehat{\sigma}_{X,Y} \leq \widehat{\sigma}_{-X,Y} \end{cases}$$

and

$$B^* = \overline{Y} -_H a^* \overline{X} \tag{8}$$

where

$$\widehat{s}_0 = \begin{cases} \infty & \text{if spr } X_i = 0 \text{ for all } i = 1, \ldots, n \\ \min\left\{\frac{\text{spr } Y_i}{\text{spr } X_i} : \text{spr } X_i \neq 0\right\} & \text{otherwise} \end{cases}$$

The basic interval linear model (6) has been criticized in some respects due to its lack of flexibility. It is straightforward to show that from (6) the following linear relationships for the *mid* and *spr* components of the intervals X and Y are transferred:

$$\operatorname{mid} Y = a \operatorname{mid} X + \operatorname{mid} \varepsilon \quad \text{and} \quad \operatorname{spr} Y = |a| \operatorname{spr} X + \operatorname{spr} \varepsilon \tag{9}$$

Since both equations involve the same (in absolute value) regression coefficient, the model is somehow restrictive for many real-life applications. With the aim of allowing more versatility in the relationship between X and Y, a new linear regression model for random intervals based on set arithmetic has been introduced in [3].

3.3 Simple Flexible Linear Model M: Blanco-Fernández et al. (2011)

The interval linear model introduced in [3] is based on the so-called *canonical decomposition* of intervals. Each $A \in \mathcal{K}_c(\mathbb{R})$ can be written as $A = \operatorname{mid} A[1 \pm 0] + \operatorname{spr} A[0 \pm 1]$. This expression allows us to work separately with the *mid* and *spr* components of A, but keeping the interval arithmetic connection. Obviously, the (inf, sup)-representation of the intervals $[1 \pm 0]$ and $[0 \pm 1]$ is $[1, 1]$ and $[-1, 1]$, respectively.

The so-called Model M between two random intervals X and Y is formalized as follows:

$$Y = \alpha \operatorname{mid} X[1 \pm 0] + \beta \operatorname{spr} X[0 \pm 1] + \gamma\,[1 \pm 0] + \varepsilon\ , \tag{10}$$

where α and β are the regression coefficients, γ is an intercept term affecting the *mid* component of Y, and ε is a random interval-valued error such that $E(\varepsilon|X) = [-\delta, \delta] \in \mathcal{K}_c(\mathbb{R})$ (with $\delta \geq 0$). To simplify notation we can define $B = [\gamma - \delta, \gamma + \delta] \in \mathcal{K}_c(\mathbb{R})$ and express the regression function associated with Model (10) as

$$E(Y|X) = \alpha X^M + \beta X^S + B, \tag{11}$$

where $X^M = \operatorname{mid} X[1 \pm 0]$ and $X^S = \operatorname{spr} X[0 \pm 1]$.

It is easy to show that from (10) the subsequent linear relationships for *mid* and *spr* variables follow:

$$\operatorname{mid} Y = \alpha \operatorname{mid} X + \gamma + \operatorname{mid} \varepsilon \quad \text{and} \quad \operatorname{spr} Y = |\beta| \operatorname{spr} X + \operatorname{spr} \varepsilon\ .$$

Since α and β are different in general, Model M is more flexible than the basic linear model in (6).

Remark 3. The particular definition of X^S allows us to assume without loss of generality that $\beta \geq 0$ (see [3]).

In [3] the LS estimation of the model (10) has been analitycally solved. Analogously to the estimation process in [7], the regression estimates are searched over a suitable feasible set which assures their coherency with the theoretical model. Thus, given a random sample $\{X_i, Y_i\}_{i=1}^n$ from (X, Y), the estimators for the parameters of the model (10) have the following expressions:

$$\widehat{\alpha} = \frac{\widehat{\sigma}_{X^M,Y}}{\widehat{\sigma}^2_{X^M}} \ ,$$

$$\widehat{\beta} = \min\left\{\widehat{s}_0 \ , \ \max\left\{0 \ , \ \frac{\widehat{\sigma}_{X^S,Y}}{\widehat{\sigma}^2_{X^S}}\right\}\right\} \ ,$$

$$\widehat{\gamma} = \mathrm{mid}\widehat{B} \tag{12}$$

and

$$\widehat{\delta} = \mathrm{spr}\,\widehat{B} \ ,$$

where $\widehat{B} = \overline{Y} -_H \left(\widehat{\alpha}\overline{X^M} + \widehat{\beta}\,\overline{X^S}\right)$.

4 Comparative Study

The practical applicability of the interval regression models revised in Section 3 is tested over the same sample data obtained from a real-life example. Moreover, the empirical behaviour of the estimated models is investigated by means of some simulation studies.

4.1 Real-life Example

The sample data set in Table 2 corresponds to a real-life case study which has been previously considered in other works to illustrate different aspects regarding regression problems for interval-valued data (see, for instance, [3], [5], [7]). Data have been supplied in 1997 by the Nephrology Unit of the Hospital *Valle del Nalón* in Asturias, Spain, to members of the *SMIRE* Research Group (http://bellman.ciencias.uniovi.es/SMIRE). They correspond to the fluctuations over a day of the systolic and the diastolic blood pressure for a sample of patients who were hospitalized in that hospital. For some purposes, physicians focus the interest on the range of variation (minimum-maximum) of these magnitudes for a patient in a day, so only the lowest and the highest values on the set of daily registers for the pressures of each patient are recorded. Therefore, these characteristics can be modelled by means of the

random intervals $X =$'fluctuation of the systolic blood pressure of a patient over a day' and $Y =$'fluctuation of the diastolic blood pressure of the patient over the same day'. If we are interested on analyzing the linear relationship between the pressure fluctuations, an interval linear model between X and Y can be estimated. From the sample data in Table 2, the estimation of the linear models presented in Sections 3.1, 3.2 and 3.3 are, respectively

$$\widehat{Y} = 0.4384X + [0.9644, 2.7586] \,, \tag{13}$$

$$\widehat{Y} = 0.4285X + [1.0695, 3.0223] \,, \tag{14}$$

$$\widehat{Y} = 0.4527X^M + 0.2641X^S + 1.6920[1 \pm 0] + [-1.5502, 1.5502] \,. \tag{15}$$

Table 2 Daily systolic (X) and diastolic (Y) blood pressure fluctuations of 59 patients

X	Y	X	Y	X	Y
[11.8,17.3]	[6.3,10.2]	[11.9,21.2]	[4.7,9.3]	[9.8,16.0]	[4.7,10.8]
[10.4,16.1]	[7.1,10.8]	[12.2,17.8]	[7.3,10.5]	[9.7,15.4]	[6.0,10.7]
[13.1,18.6]	[5.8,11.3]	[12.7,18.9]	[7.4,12.5]	[8.7,15.0]	[4.7,8.6]
[10.5,15.7]	[6.2,11.8]	[11.3,21.3]	[5.2,11.2]	[14.1,25.6]	[7.7,15.8]
[12.0,17.9]	[5.9,9.4]	[14.1,20.5]	[6.9,13.3]	[10.8,14.7]	[6.2,10.7]
[10.1,19.4]	[4.8,11.6]	[9.9,16.9]	[5.3,10.9]	[11.5,19.6]	[6.5,11.7]
[10.9,17.4]	[6.0,11.9]	[12.6,19.7]	[6.0,9.8]	[9.9,17.2]	[4.2,8.6]
[12.8,21.0]	[7.6,12.5]	[9.9,20.1]	[5.5,12.1]	[11.3,17.6]	[5.7,9.5]
[9.4,14.5]	[4.7,10.4]	[8.8,22.1]	[3.7,9.4]	[11.4,18.6]	[4.6,10.3]
[14.8,20.1]	[8.8,13.0]	[11.3,18.3]	[5.5,8.5]	[14.5,21.0]	[10.0,13.6]
[11.1,19.2]	[5.2,9.6]	[9.4,17.6]	[5.6,12.1]	[12.0,18.0]	[5.9,9.0]
[11.6,20.1]	[7.4,13.3]	[10.2,15.6]	[5.0,9.4]	[10.0,16.1]	[5.4,10.4]
[10.2,16.7]	[3.9,8.4]	[10.3,15.9]	[5.2,9.5]	[15.9,21.4]	[9.9,12.7]
[10.4,16.1]	[5.5,9.8]	[10.2,18.5]	[6.3,11.8]	[13.8,22.1]	[7.0,11.8]
[10.6,16.7]	[4.5,9.5]	[11.1,19.9]	[5.7,11.3]	[8.7,15.2]	[5.0,9.5]
[11.2,16.2]	[6.2,11.6]	[13.0,18.0]	[6.4,12.1]	[12.0,18.8]	[5.3,10.5]
[13.6,20.1]	[6.7,12.2]	[10.3,16.1]	[5.5,9.7]	[9.5,16.6]	[5.4,10.0]
[9.0,17.7]	[5.2,10.4]	[12.5,19.2]	[5.9,10.1]	[9.2,17.3]	[4.5,10.7]
[11.6,16.8]	[5.8,10.9]	[9.7,18.2]	[5.4,10.4]	[8.3,14.0]	[4.5,9.1]
[9.8,15.7]	[5.0,11.1]	[12.7,22.6]	[5.7,10.1]		

The estimation of the regression parameter a in (13) entails that some sample intervals do not keep the coherency of the interval arithmetic. For instance, for the 29-th individual $(X_{29}, Y_{29}) = ([8.8, 22.1], [3.7, 9.4])$ it is easy to verify that $\operatorname{spr} Y_{29} = 2.85 < 0.4384 \operatorname{spr} X_{29} = 2.9154$, so it would be impossible that $\operatorname{spr} Y_{29} = 0.4384 \operatorname{spr} X_{29} + e_{29}$ with $e_{29} \geq 0$. The estimate of a in (14) guarantees the existence of all the residuals and, consequently, the coherency of the estimation with the theoretical model (6). Finally, the estimated model (15) allows more flexibility to predict the *mid* and *spr* components of the diastolic blood pressure of a patient from the corresponding values of his/her systolic pressure, since they are forecasted by means of different regression parameter estimates; namely, $\operatorname{mid} \widehat{Y} = 0.4527 \operatorname{mid} X + 1.692$ and $\operatorname{spr} \widehat{Y} = 0.2641 \operatorname{spr} X + 1.5502$. The value of $MSE = \frac{1}{n} \sum_{i=1}^{n} d_\theta^2(Y_i, \widehat{Y}_i)$ of the estimated models (14) and (15) is 0.9577 and 0.9489 respectively.

4.2 Simulation Studies

The empirical behaviour of the estimation methods presented in Section 3 is checked by means of the Monte Carlo method. Let $\operatorname{mid} X, \operatorname{mid} \varepsilon \sim N(0,1)$, $\operatorname{spr} X, \operatorname{spr} \varepsilon \sim \chi^2_1$ be independent random variables. Two different theoretical linear models will be investigated in order to cover all the existing linear structures being examined. Firstly, we define Y_1 by means of a linear model in terms of X with the basic structure based on interval arithmetic:

$$Y_1 = X + \varepsilon \ , \tag{16}$$

where $a = 1$ and $B = E(\varepsilon|X) = [-1, 1]$. The regression problems developed in [5] (see Section 3.1) and [7] (see Section 3.2) solve the LS estimation of (16) by means of different techniques. Let us denote these estimation methods by $M_{(3.1)}$ and $M_{(3.2)}$, respectively. In order to compare both estimation approaches, $k = 100000$ samples of different sample sizes were simulated from the theoretical situation; in each iteration the models were estimated and, finally, the mean value and the mean square error for each estimator (denoted in general by $\widehat{\nu}$) were computed. In Table 3 the results are gathered. It is shown that the regression estimate of a obtained with method $M_{(3.1)}$ seems to be unbiased, whereas the estimate from $M_{(3.2)}$ is asymptotically unbiased. However, the mean square error of the regression estimator is greater for $M_{(3.1)}$ in all the cases. The estimation of B is similar in both methods, although it depends on the results for a.

Table 3 Comparison between the estimation methods $M_{(3.1)}$ and $M_{(3.2)}$

Parameter	n	$\widehat{E}(\widehat{\nu})$		Ratio
		$M_{(3.1)}$	$M_{(3.2)}$	$MSE_{(3.1)}/MSE_{(3.2)}$
$a = 1$	10	0.9913	0.9047	1.6696
	30	0.9978	0.9472	2.0499
	100	0.9998	0.9710	2.0599
$B = [-1, 1]$	10	[-1.0013,1.0017]	[-1.0816,1.0824]	1.0657
	30	[-0.9995,0.9996]	[-1.0508,1.0509]	1.0501
	100	[-1.0007,0.9998]	[-1.0294,1.0285]	1.0476

It is easy to check that (16) is equivalently expressed as $Y_1 = X^M + X^S + \varepsilon$, so the estimation of Model M proposed in [3] and revised in Section 3.3 can be also applied to estimate (16). Since in this case different regression parameters are estimated, they cannot be directly compared with coefficient a estimated in $M_{(3.1)}$ and $M_{(3.2)}$. The mean value and mean square error for the estimates of Model M are shown in Table 4. The estimates seem to be asymptotically unbiased, and MSE values tend to 0 as n increases.
Let Y_2 now be defined in terms of X by means of the flexible model

$$Y_2 = 3X^M + X^S + \varepsilon \ . \tag{17}$$

Table 4 Estimation of the Model M for $Y_1 = X^M + X^S + \varepsilon$

Parameter	n	$\widehat{E}(\widehat{\nu})$	$MSE(\widehat{\nu})$
$\alpha = 1$	10	1.0005	0.1441
	30	0.9978	0.0377
	100	0.9995	0.0102
$\beta = 1$	10	0.8584	0.0910
	30	0.9213	0.0199
	100	0.9594	0.0048
$B = [-1, 1]$	10	[-1.1135,1.1127]	0.2167
	30	[-1.0747,1.0697]	0.0648
	100	[-1.0401,1.0400]	0.0188

In this case, the estimation methods $M_{(3.1)}$ and $M_{(3.2)}$ are not able to capture the theoretical information about the different regression parameters relating the *mid* and *spr* values of the intervals, $\alpha = 3$ and $\beta = 1$, respectively. They provide a unique value for both estimates. Consequently, the estimation of Model M is the best (of the presented ones) technique to solve the estimation of (17). In Table 5 the empirical results for the estimation of (17) by means of $M_{(3.2)}$ and Model M are shown. In this case, the computation of $MSE(\widehat{\nu})$ for $M_{(3.2)}$ makes no sense (since there is not a unique theoretical parameter to compare with). Thus, the mean square errors of the estimated models, computed by $MSE = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n} d_\theta^2(Y_i^{(k)}, \widehat{Y}_i^{(k)})$ have been calculated. Obviously, Model M estimates the regression parameters differently, and it provides a lower mean square error in the estimation.

Table 5 Estimation methods $M_{(3.2)}$ and Model M for $Y_2 = 3X^M + X^S + \varepsilon$

Model $M_{(3.2)}$	n	$\widehat{E}(\widehat{\nu})$
a	10	1.0435
	30	1.0050
	100	1.0005
$B = [-1, 1]$	10	[-0.9559,0.9683]
	30	[-1.0030,0.9938]
	100	[-0.9938,1.0030]
MSE	10	4.9470
	30	5.4524
	100	5.6233

Model M	n	$\widehat{E}(\widehat{\nu})$
$\alpha = 3$	10	3.0002
	30	2.9977
	100	2.9999
$\beta = 1$	10	0.8587
	30	0.9220
	100	0.9594
$B = [-1, 1]$	10	[-1.1128,1.1128]
	30	[-1.0787,1.0722]
	100	[-1.0411,1.0386]
MSE	10	1.3646
	30	1.5681
	100	1.6379

5 Conclusions

Linear regression problems for interval-valued data can be addressed by means of different approaches. The natural extension from the classical real-valued case consists on formalizing a linear model between two or more ran-

dom intervals based on set arithmetic. Various models have been proposed following this idea. In this paper, several interval simple linear models based on interval arithmetic are revised, highlighting the main characteristics and the estimation process associated with each of them. From the existing models, the extension to the multivariate case may be addressed by considering several independent random intervals. Moreover, alternative models for interval-valued data based on the usual set arithmetic could be formalized.

Acknowledgements This research has been partially supported by the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02 and the Short-Term Scientific Missions associated with the COST Action IC0702 Ref. 170510-000000 and 010611-008221. Their financial support is gratefully acknowledged.

References

1. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. *Mathware Soft Comp* 2:71–84
2. Billard L, Diday E (2003) From the Statistics of data to the Statistics of knowledge: Symbolic Data Analysis. *J Amer Stat Assoc* 98:470-487
3. Blanco-Fernández A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. *Comp Stat Data Anal* 55(9):2568–2578
4. Diamond P (1990) Least squares fitting of compact set-valued data. *J Math Anal Appl* 147:531–544
5. Gil MA, Lubiano A, Montenegro M, López-García MT (2002) Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika* 56:97–111
6. González-Rodríguez G, Colubi A, Coppi R, Giordani P (2006) On the estimation of linear models with interval-valued data. *Proc. 17th IASC*. Physica-Verlag, Heidelberg
7. González-Rodríguez G, Blanco A, Corral N, Colubi A (2007) Least squares estimation of linear regression models for convex compact random sets. *Adv D Anal Class* 1:67-81
8. González-Rodríguez G, Blanco A, Colubi A, Lubiano MA (2009) Estimation of a simple linear regression model for fuzzy random variables. *Fuzzy Sets Syst* 160(3):357–370
9. Körner R, Näther W (2002) On the variance of random fuzzy variables. *Stat Mod Anal Manag Fuzzy D*, 22–39. Physica-Verlag, Heidelberg
10. Huber C, Solev V, Vonta F (2009) Interval censored and truncated data: Rate of convergence of NPMLE of the density. *J Stat Plann Infer* 139:1734–1749
11. Jahanshahloo GR, Hosseinzadeh LF, Rostamy MM (2008) A generalized model for data envelopment analysis with interval data. *Appl Math Model* 33:3237–3244
12. Lima Neto EA, DeCarvalho FAT (2010) Constrained linear regression models for symbolic interval-valued variables. *Comp Stat Data Anal* 54:333–347
13. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Inform Sci* 179(23):3964–3972
14. Wang G, Zhang Y (1992) The theory of fuzzy stochastic processes. *Fuzzy Sets Syst* 51:161–17

Bootstrap Confidence Intervals for the Parameters of a Linear Regression Model with Fuzzy Random Variables

Maria Brigida Ferraro[1], Renato Coppi[1], and Gil González-Rodríguez[2]

Abstract Confidence intervals for the parameters of a linear regression model with a fuzzy response variable and a set of real and/or fuzzy explanatory variables are investigated. The family of LR fuzzy random variables is considered and an appropriate metric is suggested for coping with this type of variables. A class of linear regression models is then proposed for the center and for suitable transformations of the spreads in order to satisfy the non-negativity conditions for the latter ones. Confidence intervals for the regression parameters are introduced and discussed. Since there are no suitable parametric sampling models for the imprecise variables, a bootstrap approach has been used. The empirical behavior of the procedure is analyzed by means of simulated data and a real-case study.

1 Introduction

The study of relationships between variables is a crucial issue in the investigation of natural and social phenomena. Of particular relevance, in this respect, is the analysis of the link between a "response" variable, say Y, and a set of "explanatory" variables, say X_1, X_2,...,X_p.

When approaching this problem from a statistical viewpoint, we realize that several sources of uncertainty may affect the analysis. These concern: a) sampling variability; b) partial or total ignorance about the kind of relationship between Y and $(X_1, ..., X_p)$; c) imprecision/vagueness in the way statistical data concerning these variables are measured (see, for instance, [2] for a detailed examination of the various sources of uncertainty in Regression Analysis).

[1] Dipartimento di Scienze Statistiche — SAPIENZA Università di Roma, Italy, {mariabrigida.ferraro,renato.coppi}@uniroma1.it
[2] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain, gil@uniovi.es

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 33–42.
springerlink.com

In the present paper we focus our attention on sources a) and c). To this purpose, the notion of Fuzzy Random Variable (FRV) is utilized [11]. By means of it, we simultaneously take into account the uncertainty due to the randomness and that pertaining to the imprecision/vagueness of the observed data. A regression model, in this context, aims at establishing a link between a random fuzzy response variable, say $\widetilde{Y}$, and a set of random fuzzy explanatory variables, say $\widetilde{X}_1$,...,$\widetilde{X}_p$. When this link is expressed in functional terms, the model may be written in the following way:

$$\widetilde{E}\left(\widetilde{Y}\right) = \widetilde{f}\left(\widetilde{X}_1, ..., \widetilde{X}_p\right),$$

where both $\widetilde{E}(\cdot)$ and $\widetilde{f}(\cdot)$ are fuzzy sets.

A complete treatment of the above model in the framework of parametric inference would require the explicitation of the family of joint densities

$$p\left(\widetilde{Y}, \widetilde{X}_1, ..., \widetilde{X}_p / \underline{\theta}\right),$$

where $\underline{\theta}$ is a vector of parameters, or at least of the conditional densities

$$p\left(\widetilde{Y} / \widetilde{X}_1, ..., \widetilde{X}_p, \underline{\theta}\right).$$

An attempt in this direction has been made by Näther and co-authors [9, 10, 12]. Unfortunately, several limitations have been found when trying to extend a complete inferential fully parametric theory of linear models to the case of fuzzy variables. One of the causes of these limitations consists in the lack of an appropriate family of sampling models for FRVs supporting the development of a complete inferential theory (consider, for example, the difficulty in defining a suitable notion of Normal FRVs).

In the present paper a different approach is adopted. First, the membership functions of the involved variables are formalized in terms of LR fuzzy numbers (in particular triangular fuzzy numbers characterized by three quantities: center, left spread, right spread). Then an appropriate distance between triangular fuzzy numbers is introduced and an isometry is found between the space of triangular fuzzy numbers and $\mathbb{R}^3$. This allows the construction of a parametric regression model linking respectively the center, left spread and right spread of the response variable to the centers and spreads of the explanatory variables. Suitable transformations of the spreads of the response variable are utilized in order to ensure the non-negativity of the estimated spreads.

While the point estimation problem concerning this model has been dealt with in previous works (see [6, 7, 8], the main objective of the present paper consists in the evaluation of the sampling variation of the estimated regression parameters. This is achieved by means of confidence intervals, which are constructed by applying an appropriate bootstrap procedure.

The work is organized as follows. In Section 2 some preliminary notions concerning fuzzy sets, LR fuzzy numbers and their distance, the definition of FRVs and their properties are briefly recalled. In Section 3 the linear regression model for LR fuzzy variables is introduced along with the procedure for estimating its parameters. The construction of bootstrap confidence intervals is illustrated in Section 4. A simulation study and a real-case analysis are described in Section 5, while concluding remarks are made in Section 6.

2 Preliminaries

Given a universe U of elements, a fuzzy set $\widetilde{A}$ is defined through the so-called *membership function* $\mu_{\widetilde{A}} : U \to [0, 1]$. For a generic $x \in U$, the membership function expresses the extent to which x belongs to $\widetilde{A}$. Such a degree ranges from 0 (complete non-membership) to 1 (complete membership).

A particular class of fuzzy sets is the LR family, whose members are the so-called *LR fuzzy numbers*. The space of the LR fuzzy numbers is denoted by $\mathcal{F}_{LR}$. A nice property of the LR family is that its elements can be determined uniquely in terms of the mapping $s : \mathcal{F}_{LR} \to \mathbb{R}^3$, i.e., $s(\widetilde{A}) = s_{\widetilde{A}} = (A^m, A^l, A^r)$. This implies that $\widetilde{A}$ can be expressed by means of three real-valued parameters, namely, the center (A^m) and the (non-negative) left and right spreads (A^l and A^r, respectively). In what follows we indistinctly use $\widetilde{A} \in \mathcal{F}_{LR}$ or (A^m, A^l, A^r). The membership function of $\widetilde{A} \in \mathcal{F}_{LR}$ can be written as

$$\mu_{\widetilde{A}}(x) = \begin{cases} L\left(\frac{A^m - x}{A^l}\right) & x \leq A^m,\ A^l > 0, \\ 1_{\{A^m\}}(x) & x \leq A^m,\ A^l = 0, \\ R\left(\frac{x - A^m}{A^r}\right) & x > A^m,\ A^r > 0, \\ 0 & x > A^m,\ A^r = 0, \end{cases} \tag{1}$$

where the functions L and R are particular decreasing shape functions from $\mathbb{R}^+$ to $[0, 1]$ such that $L(0) = R(0) = 1$ and $L(x) = R(x) = 0, \forall x \in \mathbb{R} \setminus [0, 1]$, and 1_I is the indicator function of a set I. $\widetilde{A}$ is a *triangular* fuzzy number if $L(z) = R(z) = 1 - z$, for $0 \leq z \leq 1$.

The operations considered in $\mathcal{F}_{LR}$ are the natural extensions of the Minkowski sum and the product by a positive scalar for intervals. In detail, the sum of $\widetilde{A}$ and $\widetilde{B}$ in $\mathcal{F}_{LR}$ is the LR fuzzy number $\widetilde{A} + \widetilde{B}$ so that

$$(A^m, A^l, A^r) + (B^m, B^l, B^r) = (A^m + B^m, A^l + B^l, A^r + B^r),$$

and the product of $\widetilde{A} \in \mathcal{F}_{LR}$ by a positive scalar γ is

$$\gamma(A^m, A^l, A^r) = (\gamma A^m, \gamma A^l, \gamma A^r).$$

Yang & Ko [13] defined a distance between two LR fuzzy numbers $\widetilde{X}$ and $\widetilde{Y}$ as follows

$$D^2_{LR}(\widetilde{X},\widetilde{Y}) = (X^m - Y^m)^2 + [(X^m - \lambda X^l) - (Y^m - \lambda Y^l)]^2 + [(X^m + \rho X^r) - (Y^m + \rho Y^r)]^2,$$

where the parameters $\lambda = \int_0^1 L^{-1}(\omega)\mathrm{d}\omega$ and $\rho = \int_0^1 R^{-1}(\omega)\mathrm{d}\omega$ take into account the shape of the membership function. For instance, in the triangular case, we have that $\lambda = \rho = \frac{1}{2}$. In what follows it is necessary to embed the space $\mathcal{F}_{LR}$ into $\mathbb{R}^3$ by preserving the metric. For this reason a generalization of the Yang and Ko metric can be derived [6]. For $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3) \in \mathbb{R}^3$, it is

$$D^2_{\lambda\rho}(a,b) = (a_1 - b_1)^2 + ((a_1 - \lambda a_2) - (b_1 - \lambda b_2))^2 + ((a_1 + \rho a_3) - (b_1 + \rho b_3))^2,$$

where λ, $\rho \in \mathbb{R}^+$. $D^2_{\lambda\rho}$ will be used in the sequel as a tool for quantifying errors in the regression models we are going to introduce.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. In this context, a mapping $\widetilde{X} : \Omega \to \mathcal{F}_{LR}$ is an LR FRV if the s-representation of $\widetilde{X}$, $(X^m, X^l, X^r) : \Omega \to \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ is a random vector [11]. As for non-fuzzy random variables, it is possible to determine the moments for an LR FRV. The expectation of an LR FRV $\widetilde{X}$ is the unique fuzzy set $E(\widetilde{X})$ $(\in \mathcal{F}_{LR})$ such that $(E(\widetilde{X}))_\alpha = E(X_\alpha)$ provided that $E\|\widetilde{X}\|^2_{D_{LR}} = E(X^m)^2 + E(X^m - \lambda X^l)^2 + E(X^m + \rho X^r)^2 < \infty$, where X_α is the α-level of fuzzy set $\widetilde{X}$, that is, $X_\alpha = \{x \in \mathbb{R} | \mu_{\widetilde{X}}(x) \geq \alpha\}$, for $\alpha \in (0, 1]$, and $X_0 = cl(\{x \in \mathbb{R} | \mu_{\widetilde{X}} \geq 0\})$. Moreover, on the basis of the mapping s, we can observe that $s_{E(\widetilde{X})} = (E(X^m), E(X^l), E(X^r))$.

3 A Linear Regression Model with LR Fuzzy Variables

In our previous works, [6, 7, 8], we introduced a linear regression model for imprecise information.

In the general case, an LR fuzzy response variable $\widetilde{Y}$ and p LR fuzzy explanatory variables $\widetilde{X}_1, \widetilde{X}_2, ..., \widetilde{X}_p$ observed on a simple random sample of n statistical units, $\{\widetilde{Y}_i, \widetilde{X}_{1i}, \widetilde{X}_{2i}, ..., \widetilde{X}_{pi}\}_{i=1,...,n}$, have been taken into account. We consider the shape of the membership functions as fixed, so the fuzzy response and the fuzzy explanatory variables are determined only by means of three parameters, namely the center and the left and right spreads. We faced the non-negativity constraints of the spreads of the response variable by introducing two invertible functions $g : (0, +\infty) \longrightarrow \mathbb{R}$ and $h : (0, +\infty) \longrightarrow \mathbb{R}$, in order to make the spreads assuming all the real values. In that way we didn't solve a numerical procedure, we formalized a theoretical model and we got a complete solution for the model parameters. The model is formalized as

$$\begin{cases} Y^m = \underline{X}\,\underline{a}_m^{'} + b_m + \varepsilon_m, \\ g(Y^l) = \underline{X}\,\underline{a}_l^{'} + b_l + \varepsilon_l, \\ h(Y^r) = \underline{X}\,\underline{a}_r^{'} + b_r + \varepsilon_r, \end{cases} \tag{2}$$

where $\underline{X} = (X_1^m, X_1^l, X_1^r, ..., X_p^m, X_p^l, X_p^r)$ is the row-vector of length $3p$ of all the components of the explanatory variables, ε_m, ε_l and ε_r are real-valued random variables with $E(\varepsilon_m|\underline{X}) = E(\varepsilon_l|\underline{X}) = E(\varepsilon_r|\underline{X}) = 0$, $\underline{a}_m = (a_{mm}^1, a_{ml}^1, a_{mr}^1, ..., a_{mm}^p, a_{ml}^p, a_{mr}^p)$, $\underline{a}_l = (a_{lm}^1, a_{ll}^1, a_{lr}^1, ..., a_{lm}^p, a_{ll}^p, a_{lr}^p)$ and $\underline{a}_r = (a_{rm}^1, a_{rl}^1, a_{rr}^1, ..., a_{rm}^p, a_{rl}^p, a_{rr}^p)$ are row-vectors of length $3p$ of the parameters related to $\underline{X}$. The generic $a_{jj'}^t$ is the regression coefficient between the component $j \in \{m, l, r\}$ of $\widetilde{Y}$ (where m, l and r refer to the center Y^m and the transformations of the spreads $g(Y^l)$ and $h(Y^r)$, respectively) and the component $j' \in \{m, l, r\}$ of the explanatory variables $\widetilde{X}^t$, $t = 1, ..., p$, (where m, l and r refer to the corresponding center, left spread and right spread). For example, a_{ml}^2 represents the relationship between the center of the response, Y^m, and the left spread of the explanatory variable $\widetilde{X}^2$ (X_2^l). Finally, b_m, b_l, b_r denote the intercepts. Therefore, by means of (2), we aim at studying the relationship between the response and the explanatory variables taking into account not only the randomness due to the data generation process, but also the information provided by the spreads of the explanatory variables (the imprecision of the data), which are usually arbitrarily ignored. The covariance matrix of $\underline{X}$ is denoted by $\Sigma_{\underline{X}} = E[(\underline{X} - E\underline{X})^{'}(\underline{X} - E\underline{X})]$ and Σ stands for the covariance matrix of $(\varepsilon_m, \varepsilon_l, \varepsilon_r)$, with variances, $\sigma_{\varepsilon_m}^2$, $\sigma_{\varepsilon_l}^2$ and $\sigma_{\varepsilon_r}^2$, strictly positive and finite.

3.1 Estimation Problem

The estimation problem of the regression parameters is faced by means of the Least Squares (LS) criterion. Accordingly, parameters of model (2) are estimated by minimizing the sum of the squared distances between the observed and theoretical values of the response variable. However, as already noted, suitable transformations of the spreads are considered in (2). This allows us to use the generalized metric $D_{\lambda\rho}^2$ in the objective function of the problem. Therefore, the LS problem consists in looking for $\widehat{\underline{a}}_m$, $\widehat{\underline{a}}_l$, $\widehat{\underline{a}}_r$, $\widehat{b}_m$, $\widehat{b}_l$ and $\widehat{b}_r$ which minimize

$$\begin{aligned} \Delta_{\lambda\rho}^2 &= D_{\lambda\rho}^2((\underline{Y}^m, g(\underline{Y}^l), h(\underline{Y}^r)), ((\underline{Y}^m)^*, g(\underline{Y}^l)^*, h(\underline{Y}^r)^*)) \\ &= \sum_{i=1}^{n} D_{\lambda\rho}^2((Y_i^m, g(Y_i^l), h(Y_i^r)), ((Y_i^m)^*, g(Y_i^l)^*, h(Y_i^r)^*)), \end{aligned} \tag{3}$$

where $\underline{Y}^m$, $g(\underline{Y}^l)$ and $h(\underline{Y}^r)$ are the $n \times 1$ vectors of the observed values and $(\underline{Y}^m)^* = \mathbf{X}\underline{a}_m^{'} + \underline{1}b_m$, $g(\underline{Y}^l)^* = \mathbf{X}\underline{a}_l^{'} + \underline{1}b_l$ and $h(\underline{Y}^r)^* = \mathbf{X}\underline{a}_r^{'} + \underline{1}b_r$ are

the theoretical ones being $\mathbf{X} = (\underline{X}_1, \underline{X}_2, ..., \underline{X}_n)^{'}$ the $n \times 3p$ matrix of the explanatory variables. When estimating the regression parameters using this least squares criterion we obtain the following solution

$$\begin{array}{lll} \widehat{\underline{a}}_m^{'} = (\mathbf{X}^{c^{'}}\mathbf{X}^c)^{-1}\mathbf{X}^{c^{'}}\underline{Y}^{mc}, & & \widehat{b}_m = \overline{Y^m} - \overline{\underline{X}}\widehat{\underline{a}}_m^{'}, \\ \widehat{\underline{a}}_l^{'} = (\mathbf{X}^{c^{'}}\mathbf{X}^c)^{-1}\mathbf{X}^{c^{'}}g(\underline{Y}^l)^c, & & \widehat{b}_l = \overline{g(Y^l)} - \overline{\underline{X}}\widehat{\underline{a}}_l^{'}, \\ \widehat{\underline{a}}_r^{'} = (\mathbf{X}^{c^{'}}\mathbf{X}^c)^{-1}\mathbf{X}^{c^{'}}h(\underline{Y}^r)^c, & & \widehat{b}_r = \overline{h(Y^r)} - \overline{\underline{X}}\widehat{\underline{a}}_r^{'}, \end{array}$$

where

$$\begin{array}{c} \underline{Y}^{mc} = \underline{Y}^m - \underline{1}\overline{Y^m}, \\ g(\underline{Y}^l)^c = g(\underline{Y}^l) - \underline{1}\overline{g(Y^l)}, \\ h(\underline{Y}^r)^c = h(\underline{Y}^r) - \underline{1}\overline{h(Y^r)} \end{array}$$

are the centered values of the response variables,

$$\mathbf{X}^c = \mathbf{X} - \underline{1}\,\overline{\underline{X}}$$

is the centered matrix of the explanatory variables and, $\overline{Y^m}$, $\overline{g(Y^l)}$, $\overline{h(Y^r)}$ and $\overline{\underline{X}}$ denote, respectively, the sample means of Y^m, $g(Y^l)$, $h(Y^r)$ and $\underline{X}$. The Yang and Ko metric involves differences between re-scaled intervals. Since the same parameters λ and ρ are considered for both the observed and the predicted values, the solution of the minimization problem does not depend on the values of these parameters.

4 Bootstrap Confidence Intervals

As in classical Statistics, in this case it is useful to estimate the regression parameters not only by a single value but by a confidence interval too. These intervals represent the reliability of the estimates. How likely the interval is to contain the parameter is determined by the confidence level $1 - \alpha$.
Since there are no realistic models in the context of FRVs, we introduce a bootstrap approach. In literature, there exist different bootstrap approaches to construct confidence intervals in a real-valued variables context. In this work, we consider confidence intervals based on bootstrap percentiles (for more details, see [5, 1]).
We consider B bootstrap samples drawn with replacement from the observed sample $\{\widetilde{Y}_i, \widetilde{X}_{1i}, \widetilde{X}_{2i}, ..., \widetilde{X}_{pi}\}_{i=1,...,n}$. For each sample we compute the estimators of the regression parameters. In this way we obtain sequences of B bootstrap estimators, that represent the empirical distributions of the estimators. Let $\widehat{F}$ be the cumulative distribution function of the bootstrap replications of each estimator. The $1 - \alpha$ percentile interval is defined by means of the percentiles of $\widehat{F}$. For example, for the estimator of a_{ml}^1, $\widehat{F}^{-1}(\alpha/2)$ is

equal to $\widehat{a}_{ml}^{1^{*(\alpha/2)}}$, that is, the $100 \cdot (\alpha/2)$th percentile of the bootstrap distribution. In details, $\widehat{a}_{ml}^{1^{*(\alpha/2)}}$ is the $B \cdot (\alpha/2)$th value in the ordered list of the B bootstrap estimators $\left(\widehat{a}_{ml}^{1^{*(1)}}, \widehat{a}_{ml}^{1^{*(2)}}, \ldots, \widehat{a}_{ml}^{1^{*(B)}}\right)$. The bootstrap percentile interval for a_{ml}^1 is defined as:

$$CI_P(a_{ml}^1) \;=\; \left[\widehat{F}^{-1}(\alpha/2), \widehat{F}^{-1}(1-\alpha/2)\right] \;=\; \left[\widehat{a}_{ml}^{1^{*(\alpha/2)}}, \widehat{a}_{ml}^{1^{*(1-\alpha/2)}}\right]$$

The bootstrap percentile confidence interval for a_{ml}^1 is obtained by means of the following algorithm

Algorithm

Step 1: Draw a sample of size n with replacement

$$\left\{(\underline{X}_i^*, Y_i^{m*}, Y_i^{l^*}, Y_i^{r^*})\right\}_{i=1,\ldots,n},$$

from the original sample $\left\{(\underline{X}_i, Y_i^m, Y_i^l, Y_i^r)\right\}_{i=1,\ldots,n}$.

Step 2: Compute the bootstrap estimate $\widehat{a}_{ml}^{1^*}$.

Step 3: Repeat Steps 1 and 2 B times to get different sets of estimators for the regression parameter.

Step 4: Approximate the lower and upper limits of the interval by the quantiles of the empirical distribution obtained at Step3. That is, the values in position $[(\alpha/2)B]+1$ and $[(1-\alpha/2)B]$ of the ordered empirical distribution. We indicate those values as $\widehat{a}_{ml_L}^{1^*}$ and $\widehat{a}_{ml_U}^{1^*}$. Thus the percentile confidence interval for a_{ml}^1 at the confidence level $1-\alpha$ is

$$CI_P(a_{ml}^1) = \left[\widehat{a}_{ml_L}^{1^*}, \widehat{a}_{ml_U}^{1^*}\right]$$

An analogous algorithm could be used to construct the bootstrap percentile confidence intervals for all the regression parameters.

5 Empirical Results

In order to check the empirical behaviour of the bootstrap approach to construct confidence intervals for the regression parameters, some simulation studies and a real-case example have been developed.

5.1 Simulation Studies

We consider a theoretical situation in which an LR fuzzy response $\widetilde{Y}$, an LR fuzzy explanatory variable $\widetilde{X}_1$ and a real explanatory variable X_2 are taken into account. We deal with the following independent real random variables: X_1^m behaving as $Norm(0,1)$ random variable, X_1^l and X_1^r as χ_1^2 and χ_2^2, respectively, X_2 as $U(-2,2)$, ε_m as $Norm(0,1)$, ε_l and ε_r as $Norm(0,0.5)$. The response variables are constructed in the following way:

$$\begin{cases} Y^m = 2X_1^m + 0.5X_1^l + 0.4X_1^r + X_2 + \varepsilon_m, \\ Y^2 = g(Y^l) = -1X_1^m + 0.3X_1^l - 0.4X_1^r + 2X_2 + +\varepsilon_l, \\ Y^3 = h(Y^r) = 1.2X_1^m + X_1^l - 0.7X_1^r - X_2 + +\varepsilon_r, \end{cases}$$

During the experiment we employ $B = 1000$ replications of the bootstrap estimator and we carry out $N = 10.000$ simulations with the confidence level $1-\alpha = 0.95$ for different sample sizes ($n = 30, 50, 100, 200, 300$). We compute the empirical confidence levels as the proportion of bootstrap confidence intervals that include the theoretical parameter (on N). The empirical values are reported in Table 1. Since the values gathered in Table 1 tend to the nominal confidence level, as n increases, we can conclude that the bootstrap algorithm perform well in this context.

Table 1 Empirical confidence level of the bootstrap CIs for the regression parameters.

n	30	50	100	200	300
$CI(a^1_{mm})$	.9440	.9352	.9410	.9390	.9475
$CI(a^1_{ml})$	.9381	.9378	.9351	.9382	.9443
$CI(a^1_{mr})$	.9384	.9392	.9410	.9408	.9463
$CI(a^2_m)$	.9408	.9411	.9431	.9469	.9464
$CI(a^1_{lm})$	.9427	.9348	.9407	.9429	.9484
$CI(a^1_{ll})$	.9363	.9377	.9330	.9394	.9444
$CI(a^1_{lr})$	.9361	.9341	.9400	.9410	.9413
$CI(a^2_l)$	.9357	.9397	.9551	.9485	.9489
$CI(a^1_{rm})$	.9401	.9364	.9383	.9466	.9466
$CI(a^1_{rl})$	.9371	.9324	.9344	.9405	.9404
$CI(a^1_{rr})$	.9375	.9383	.9403	.9425	.9430
$CI(a^2_r)$	.9365	.9479	.9450	.9456	.9457
$CI(b_m)$	.9444	.9405	.9467	.9450	.9517
$CI(b_l)$	.9424	.9421	.9475	.9479	.9486
$CI(b_r)$	.9409	.9435	.9471	.9469	.9453

5.2 A Real-case Study

We consider the students' satisfaction of a course. In order to evaluate it, their subjective judgments/ perceptions are observed on a sample of $n = 64$ students (see, for more details, [8]). For any student, four characteristics are observed: the overall assessment of the course, the assessment of the teaching staff, the assessment of the course content and the average mark (single-valued variable). We managed them in terms of fuzzy variables, in particular of triangular type (hence $\lambda = \rho = 1/2$). For analyzing the linear dependence of the overall assessment of the course ($\widetilde{Y}$) on the assessment of the teaching staff ($\widetilde{X}_1$), the assessment of the course contents ($\widetilde{X}_2$) and the average mark (X_3), the proposed linear regression model is employed based on a sample of 64 students. In order to overcome the problem about the non-negativity of spread estimates, we fix the logarithmic transformation (that is, $g = h = \ln$). Through the LS procedure we obtain the following estimated model

$$\begin{cases} \widehat{Y^m} = 1.08X_1^m + 0.13X_1^l - 0.07X_1^r \\ \quad -0.17X_2^m - 0.89X_2^l + 0.66X_2^r - 1.12X_3 + 34.06 \\ \widehat{Y^l} = exp(0.01X_1^m + 0.02X_1^l + 0.02X_1^r \\ \quad +0.00X_2^m + 0.03X_2^l + 0.01X_2^r - 0.00X_3 + 0.67) \\ \widehat{Y^r} = exp(0.00X_1^m + 0.03X_1^l - 0.02X_1^r \\ \quad -0.01X_2^m + 0.03X_2^l + 0.01X_2^r + 0.04X_3 + 1.01) \end{cases}$$

For each one of the regression parameters we obtain the bootstrap percentile confidence intervals reported in Table 2.

Table 2 Bootstrap percentile CIs for the regression parameters at a confidence level equal to 0.95.

$CI(a^1_{mm})$	$[.7888, 1.3403]$	$CI(a^1_{lm})$	$[-.0018, .0199]$	$CI(a^1_{rm})$	$[-.0086, .0142]$
$CI(a^1_{ml})$	$[-.6060, .8087]$	$CI(a^1_{ll})$	$[-.0314, .0556]$	$CI(a^1_{rl})$	$[-.0161, .0633]$
$CI(a^1_{mr})$	$[-.4848, .5013]$	$CI(a^1_{lr})$	$[-.0052, .0358]$	$CI(a^1_{rr})$	$[-.0487, .0101]$
$CI(a^2_{mm})$	$[-.2878, .0324]$	$CI(a^2_{lm})$	$[-.0069, .0071]$	$CI(a^2_{rm})$	$[-.0154, -.0004]$
$CI(a^2_{ml})$	$[-1.4890, -.4884]$	$CI(a^2_{ll})$	$[.0092, .0579]$	$CI(a^2_{rl})$	$[.0021, .0688]$
$CI(a^2_{mr})$	$[.3474, .9626]$	$CI(a^2_{lr})$	$[-.0021, .0249]$	$CI(a^2_{rr})$	$[-.0002, .0330]$
$CI(a^3_m)$	$[-4.3814, .4688]$	$CI(a^3_l)$	$[-.0962, .0953]$	$CI(a^3_r)$	$[-.0473, .1964]$
$CI(b_m)$	$[5.1405, 121.9076]$	$CI(b_l)$	$[-2.0829, 3.16179]$	$CI(b_r)$	$[-3.7414, 3.5874]$

It could be noted from Table 2 that the parameters that are significant are the same as those obtained by means of a bootstrap test on the regression parameters in [8]. In detail, these are: a^1_{mm}, a^2_{ml}, a^2_{mr}, a^2_{ll}, a^2_{rm} and a^2_{rl}.

6 Concluding Remarks

In this paper a linear regression model for LR fuzzy variables has been addressed. Along with the least squares estimators, confidence intervals have been introduced and discussed. The results obtained by means of a bootstrap approach are those expected in this context. In detail, a bootstrap algorithm to approximate the bootstrap percentile confidence intervals of the parameters has been described and employed to simulated and real data.

Acknowledgements The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grant MTM2009-09440-C02-02 and the COST Action IC0702. Their financial support is gratefully acknowledged.

References

1. Blanco A, Corral N, González-Rodríguez G, Palacio A (2010) On some confidence regions to estimate a linear regression model for interval data. In: Borgelt C, González-Rodríguez G, Trutschnig W, Lubiano MA, Gil MA, Grzegorzewski P, Hryniewicz O (eds.): *Combining soft computing and statistical methods in data analysis, Advances in Intelligent and Soft Computing* 77:33–40
2. Coppi R (2008) Management of uncertainty in statistical reasoning: the case of regression analysis. *Int. J. Approx. Reason.* 47:284–305
3. Coppi R, D'Urso P, Giordani P, Santoro A (2006) Least squares estimation of a linear regression model with LR fuzzy response. *Comput. Statist. Data Anal.* 51:267–286
4. Coppi R, Gil MA, Kiers HAL (2006) The fuzzy approach to statistical analysis. *Comput. Statist. Data Anal.* 51:1–14
5. Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap.* Chapman & Hall, New York
6. Ferraro MB, Coppi R, Gonzalez-Rodriguez G, Colubi A (2010) A linear regression model for imprecise response. *Int. J. Approx. Reason.* 51:759–770
7. Ferraro MB, Colubi A, Gonzalez-Rodriguez G, Coppi R (2011) A determination coefficient for a linear regression model with imprecise response. *Environmetrics* 22:487–596
8. Ferraro MB, Giordani P (2011) A multiple linear regression model for LR fuzzy random variables. *Metrika* doi:10.1007/s00184-011-0367-3 (in press)
9. Näther W (2000) On random fuzzy variables of second order and their application to linear statistical inference with fuzzy data. *Metrika* 51:201–221
10. Näther, W (2006) Regression with fuzzy random data. *Comp. Stat. Data Anal.* 51:235–252
11. Puri ML, Ralescu DA (1986) Fuzzy random variables. *J. Math. Anal. Appl.* 114:409–422
12. Wünsche A, Näther W (2002) Least-squares fuzzy regression with fuzzy random variables. *Fuzzy Sets Syst.* 130:43–50
13. Yang MS, Ko CH (1996) On a class of fuzzy c-numbers clustering procedures for fuzzy data. *Fuzzy Sets Syst.* 84:49–60
14. Zadeh LA (1965) Fuzzy sets. *Inf. Control* 8:338–353

On the Estimation of the Regression Model M for Interval Data

Marta García-Bárzana[1], Ana Colubi[1], and Erricos J. Kontoghiorghes[2]

Abstract A linear regression model for interval data based on the natural interval-arithmetic has recently been proposed. Interval data can be identified with 2-dimensional points in $\mathbb{R} \times \mathbb{R}^+$, since they can be parametrized by its mid-point and its semi-amplitude or spread, which is non-negative. The model accounts separately for the contribution of the mid-points and the spreads through a single equation. The least squares estimation becomes a quadratic optimization problem subject to linear constraints, which guarantee the existence of the residuals. Several estimators are discussed. Namely, a closed-form estimator, the restricted least-squares estimator, an empirical estimator and an estimator based on separate models for mids and spreads have been investigated. Real-life examples are considered. Simulations are performed in order to assess the consistency and the bias of the estimators. Results indicate that the numerical and the closed-form estimator are appropriate in most of cases, while the empirical estimator and the one based on separate models are not always suitable.

1 Introduction

Often experimental researches involves non-perfect data, as missing data, or censored data. In particular, closed and bounded real-valued sets in $\mathbb{R}^p$ are useful to model information which also representing linguistic descriptions, fluctuations, grouped data images, to name but a few. Interval data are a specific case of this kind of elements. The study of linear regression models working with interval-valued variables has been addressed *mainly* by two ways:

[1] Department of Statistics and Operational Research, University of Oviedo, Spain, garciabarmarta.uo@uniovi.es · colubi@uniovi.es
[2] Department of Commerce, Finance and Shipping, Cyprus University of Technology, Cyprus, erricos@cut.ac.cy

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 43–52.
springerlink.com

(a) in terms of the separate models involving some interval components (as the midpoint and the range or the minimum and the maximun) (see Billard and Diday, 2003; Lima Neto *et al.*, 2005 and references therein) which most of the times work with symbolic interval variables; and (b) in terms of arithmetic set-based unified models (as in Diamond 1990, Gil *et al.* 2001, 2002, 2007, González-Rodríguez *et al.* 2007, Blanco-Fernández *et al.* 2011, among others). The main difference between both views is that the first approach usually fits the separate models by numerical or classical tools, but without the usual probabilistic assumptions for the regression model. This provides good fittings but non-obvious easy ways of making inferences. On the other hand, the second approach provides a natural framework to develop inferences, although the least squares approach becomes a minimization problem with strong constraints.

In Blanco-Fernández *et al.* (2011) a flexible simple linear regression model was introduced, the so-called Model M. This model is *flexible* in the sense that it accounts for relationship between mid points and the radius of the involved random intervals. A comparison of several regression estimators of Model M will be addressed.

The rest of the paper is organized as follows: in Section 2 some preliminary about the Model M will be introduced. In Section 3 four estimation approaches of Model M will be described. In Section 4 a real-life example is analyzed to compare the behaviour of the estimators. Finally, Section 5 cointains some conclusions.

2 The Model M for Random Intervals

Hereafter, the intervals that will be considered are elements in the space $\mathcal{K}_c(\mathbb{R}) = \{[a_1, a_2] : a_1, a_2 \in \mathbb{R}, a_1 \leq a_2\}$. An interval $A \in \mathcal{K}_c(\mathbb{R})$ can be expressed in terms of its minimun and maximun or in terms or its middle point (mid) and the radius (spr). The second characterization is more usual in regression studies, as it involves non-negativity constraints which are easier to handle than the order contraints involved in the first characterization.
There is another representation for the intervals which will be used, namely, the *canonical decomposition*, defined as $A = \operatorname{mid} A\,[1 \pm 0] + \operatorname{spr} A\,[0 \pm 1]$ (see Blanco-Fernández *et al.*, 2011).

The arithmetics which will be used are the *Minkowski addition* $A + B = \{a + b : a \in A, b \in B\}$ and the product by scalars $\lambda A = \{\lambda a : a \in A\}$, with $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. The space $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear as the existence of the symmetric element with respect to the addition is not guaranteed in general, in the sense that $A + (-A) \neq \{0\}$ unless A is a singleton. A new concept of difference agreeing with the natural dif-

ference, the so-called *Hukuhara difference*, is introduced. It is defined as $A -_H B = [\inf A - \inf B, \sup A - \sup B]$ if and only if $\operatorname{spr} B \leq \operatorname{spr} A$.

Remark: If $\operatorname{spr} B > \operatorname{spr} A$, then the Hukuhara difference does not exist.

The distance used is the so-called d_τ (see Trutschnig *et al.*, 2009) defined as

$$d_\tau(A, B) = \sqrt{(1-\tau)(\operatorname{mid} A - \operatorname{mid} B)^2 + \tau\,(\operatorname{spr} A - \operatorname{spr} B)^2}$$

for all $A, B \in \mathcal{K}_c(\mathbb{R})$.
Random intervals emerged as a generalization of the real-valued random variables. Then, $\boldsymbol{y}$ is a random interval if it is $\mathcal{B}_{d_\tau}|\mathcal{A}$ measurable, being $\mathcal{B}_{d_\tau}$ the Borel σ-algebra and $\mathcal{A}$ the σ-algebra of the probabilistic space $(\Omega, \mathcal{A}, P)$.

Notation: Random intervals will be denoted with boldlowercase letters $(\boldsymbol{x})$, vectors with lowercase letters (x) and matrices with uppercase letters (X). The (Aumann) expect value is defined as $E(\boldsymbol{x}) = [E(\operatorname{mid}\boldsymbol{x}) \pm E(\operatorname{spr}\boldsymbol{x})]$, whenever $\operatorname{mid}\boldsymbol{x}$ and $\operatorname{spr}\boldsymbol{x} \in L^1(\Omega, \mathcal{A}, P)$. The Aumann expectation fulfils Fréchet principle and the Fréchet variance associated with this expectation is defined as

$$Var_\tau(\boldsymbol{x}) = \sigma^2_{\boldsymbol{x},\tau} = E(d_\tau(\boldsymbol{x}, E(\boldsymbol{x}))) = (1-\tau)\,\sigma^2_{\operatorname{mid}\boldsymbol{x}} + \tau\,\sigma^2_{\operatorname{spr}\boldsymbol{x}}$$

whenever $\operatorname{mid}\boldsymbol{x}$ and $\operatorname{spr}\boldsymbol{x}$ are integrably bounded.
As $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not a linear space, the covariance cannot be defined by mimiching the usual expression involving the arithmetic in $\mathcal{K}_c(\mathbb{R})$. However, it can be defined in $\mathbb{R}^2$ and we get the following expression

$$Cov_\tau(\boldsymbol{x}, \boldsymbol{y}) = \sigma_{\boldsymbol{x},\boldsymbol{y}} = (1-\tau)\sigma_{\operatorname{mid}\boldsymbol{x},\operatorname{mid}\boldsymbol{y}} + \tau\,\sigma_{\operatorname{spr}\boldsymbol{x},\operatorname{spr}\boldsymbol{y}}$$

whenever $\|\operatorname{mid}\boldsymbol{x}\|^2_\tau, \|\operatorname{mid}\boldsymbol{y}\|^2_\tau, \|\operatorname{spr}\boldsymbol{x}\|^2_\tau, \|\operatorname{spr}\boldsymbol{y}\|^2_\tau \in L^1(\Omega, \mathcal{A}, P)$.

Model M will relate a response random interval $\boldsymbol{y} : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ with an explanatory random interval $\boldsymbol{x} : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ as follows

$$\boldsymbol{y} = \boldsymbol{x}^M\,\alpha_1 + \boldsymbol{x}^S\,\alpha_2 + \varepsilon \tag{1}$$

where $\boldsymbol{x}^M = \operatorname{mid}\boldsymbol{x}[1\pm 0] = [\operatorname{mid}\boldsymbol{x}, \operatorname{mid}\boldsymbol{x}]$, $\boldsymbol{x}^S = \operatorname{spr}\boldsymbol{x}[0\pm 1] = [-\operatorname{spr}\boldsymbol{x}, \operatorname{spr}\boldsymbol{x}]$, α_1, α_2 and $\varepsilon \in \mathcal{K}_c(\mathbb{R})$ (see Blanco-Fernández et al, 2011).
The Model can be written in the matricial way as

$$\boldsymbol{y} = x^{Bl}\,b_\alpha + \varepsilon \tag{2}$$

with $x^{Bl} = (\boldsymbol{x}^M \,|\, \boldsymbol{x}^S) \in \mathcal{K}_c(\mathbb{R})^{1\times 2}$, $b_\alpha = (\alpha_1 \,|\, \alpha_2)^t \in \mathbb{R}^{2\times 1}$ and
$\varepsilon : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ being a random interval such that $E(\varepsilon|\boldsymbol{x}) = \Delta \in \mathcal{K}_c(\mathbb{R})$.

Remark: A property of this model is that it is not identifiable due to the fact that $\boldsymbol{x}^S = -\boldsymbol{x}^S$. However, the coefficient α_2 can be considered, without loss of generality, a non-negative vector in $\mathbb{R}$ and the space in which the

solutions to the estimation problem are, can be restricted to $\mathbb{R}^+$. In this way, the model is identifiable.

Model M entails the following separate models

$$\begin{aligned} \operatorname{mid} \boldsymbol{y} &= \alpha_1 \left(\operatorname{mid} \boldsymbol{x}\right) + \operatorname{mid} \varepsilon \\ \operatorname{spr} \boldsymbol{y} &= |\alpha_2| \left(\operatorname{spr} \boldsymbol{x}\right) + \operatorname{spr} \varepsilon. \end{aligned} \tag{3}$$

Remark: By the assumption that α_2 can be considered non-negative, the second expression can be written as

$$\operatorname{spr} \boldsymbol{y} = \alpha_2 (\operatorname{spr} \boldsymbol{x}) + \operatorname{spr} \varepsilon.$$

Thus, it is feasible to consider the estimation of α_1 and α_2 through the estimation of the separate models.

3 Estimation of the Model M

Four estimators of the regression coefficients will be considered. The first one based on the fitting of the separate models introduced in (3). Separate models have already considered to relate interval-valued variables (see Lima Neto & Carvalho 2010 among others). In this case the proposed separate models are:

$$\operatorname{mid} \boldsymbol{y} = \boldsymbol{x^c}\, b^m + \varepsilon^m \tag{4}$$

$$\operatorname{spr} \boldsymbol{y} = \boldsymbol{x^s}\, b^s + \varepsilon^s, \tag{5}$$

where $\boldsymbol{x^c} = (1, \operatorname{mid} \boldsymbol{x})$ and $\boldsymbol{x^s} = (1, \operatorname{spr} \boldsymbol{x}) \in \mathcal{K}_c(\mathbb{R})^{1\times 2}$, b^m and $b^s \in \mathbb{R}^{2\times 1}$, $\boldsymbol{y} \in \mathcal{K}_c(\mathbb{R})$ and ε^m, $\varepsilon^s \in \mathbb{R}$. Lima Neto & Carvalho impose the condition that $b^s \geq 0$ to avoid spreads ill-defined. However, b^m has no constraint to be fulfilled.

Then, let $\{\boldsymbol{y_i}, \boldsymbol{x_i}\}_{i=1,\ldots,n}$ be a random simple sample of intervals, the estimator of b^m will be:

$$\widehat{b^m} = \left[(x^c)^t \, (x^c)\right]^{-1} (x^c)^t \operatorname{mid} y \tag{6}$$

where $\operatorname{mid} y \in \mathbb{R}^{n\times 1}$ and

$$x^c = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \operatorname{mid} \boldsymbol{x_1} & \operatorname{mid} \boldsymbol{x_2} & \ldots & \operatorname{mid} \boldsymbol{x_n} \end{pmatrix} \in \mathbb{R}^{n\times 2}.$$

Parameter b^s is estimated according to Lawson and Hanson algorithm (see Lawson and Hanson, 1974) for constrained LS problems. Then the estimator of both parameters will be denoted by $\widehat{b_{sep}} = (\widehat{b^m}, \widehat{b^s})$.

Remark: The main drawback of using the separate models to estimate the coefficients is that (5) is not a linear model, due to the non-negativity constraint of the variables. Additionally, the linear independence between

the residuals and the independent variables implies further restrictions on the residuals. Thus, inferences are not straight-forward deduced.

It is possible to obtain another estimator of b_α by using sample moments. Hence, it is introduced the following proposition:

Proposition 1. *Given the random interval* $\boldsymbol{y}$ *and the vector of random intervals* x^{bl} *in the conditions of the Model* M*, the coefficients' vector* b_α *can be expressed by:*

$$b_\alpha = Cov_\tau(\boldsymbol{y}, x^{bl}) Cov_\tau(x^{bl}, x^{bl})^{-1}.$$

According to Proposition 1, an empirical estimator could be proposed based on the sample moments, namely:

$$\widehat{b_{emp}} = Cov_\tau(y, X^{bl})\, Cov_\tau(X^{bl}, X^{bl})^{-1} \tag{7}$$

with $X^{bl} \in \mathcal{K}_c(\mathbb{R})^{2\times n}$ and $y \in \mathcal{K}_c(\mathbb{R})^{1\times n}$.

The least squares estimation of b_α and the parameter Δ will be carried out from the information provided by the simple random sample of random intervals $\{\boldsymbol{y_i}, \boldsymbol{x_i}\}_{i=1,\ldots,n}$ obtained from the model:

$$y = X^{bl}\, \widehat{b_\alpha} + \hat{\varepsilon} \tag{8}$$

being

$$X^{bl} = (x^M \mid x^S) \in \mathcal{K}_c(\mathbb{R})^{n\times 2}$$

and

$$\widehat{b_\alpha} = (\hat{\alpha_1} \mid \hat{\alpha_2})^t \in \mathbb{R}^{2\times 1}.$$

It is neccesary to assure the existence of the residuals, or in other words, that the Hukuhara's difference $y -_H (X^{bl}\, \widehat{b_\alpha})$ exists. Then the expression of the constraints is:

$$\operatorname{spr}(\hat{\alpha_1}\, x^M + \hat{\alpha_2}\, x^S) \leq \operatorname{spr} y$$

which is equivalent to

$$sign(\hat{\alpha_2}) \circ |\alpha_2| \operatorname{spr} x \leq \operatorname{spr} y \equiv \hat{\alpha_2} \operatorname{spr} x \leq \operatorname{spr} y.$$

In order to assure the existance of the residuals, the least squares problem will be written as a minimization problem with linear constraints. Specifically, the aim will be to find feasible estimates of b_α and Δ minimizing the not explained variability, that is,

$$\min_{c_2 \in \Gamma} d_\tau^2(y, X^{bl}\, c + 1\Delta) \tag{9}$$

where $c = (c_1, c_2)^t \in \mathbb{R}^{2\times 1}$ and $\Gamma = \{c_2 \in [0, \infty)/c_2 \operatorname{spr} x \leq \operatorname{spr} y\}$.

Introducing the following notation, the minimization problem (9) will be transcribed into another one with some useful properties.

$$v_m = \operatorname{mid} y - \overline{\operatorname{mid} \boldsymbol{y}}1 \ ; \ F_m = \operatorname{mid} X^{bl} - (\overline{\operatorname{mid} X^{bl}})1 \qquad (10)$$
$$v_s = \operatorname{spr} y - \overline{\operatorname{spr} \boldsymbol{y}}1 \ ; \ F_s = \operatorname{spr} X^{bl} - (\overline{\operatorname{spr} X^{bl}})1,$$

where $v_m, v_s \in \mathbb{R}^{n\times 1}$ and $F_m, F_s \in \mathbb{R}^{n\times 2}$. Then, the minimization problem can be written as:

$$\min_{c_2 \in \Gamma} (1-\tau)(v_m - F_m\, c)^t (v_m - F_m\, c) + \tau\, (v_s - F_s\, c)^t (v_s - F_s\, c). \qquad (11)$$

Two possible ways of solving the problem have been proposed. The first one results in a numerical estimator and the second one in an exact expression. Concerning the first approach, as the objective function is a quadratic function and Γ is a set of linear constraints, Karush-Kuhn-Tucker (KKT) Theorem assures the existence of solution and by means of the numerical estimator, which will be denoted in the sequel by $\widehat{b_{kkt}}$, an estimation of the solution will be obtained.
On the other hand, a closed expression to estimate the regression coefficients has been obtained in Blanco-Fernández *et al.* (2011). It is given in the following proposition and will be the last one to be compared later on.

Proposition 2. *Under the conditions of Model M, the LS regression coefficients estimator is* $\widehat{b_{exact}} = (\hat{\alpha_1}, \hat{\alpha_2})$*, where:*

$$\hat{\alpha_1} = \frac{Cov(\boldsymbol{x}^M, \boldsymbol{y})}{Var(\boldsymbol{x}^M)}$$
$$\hat{\alpha_2} = \min\left\{\hat{a^0}, \max\left\{0, \frac{Cov(\boldsymbol{x}^S, \boldsymbol{y})}{Var(\boldsymbol{x}^S)}\right\}\right\}$$

being $\hat{a^0} = \min\left\{\frac{\operatorname{spr} y_i}{\operatorname{spr} x_i}\right\}\ \forall i \in \{1, \ldots, n\}$.

According to Blanco-Fernández *et al.* (2011), given $\widehat{b_\alpha}$ any estimator of b_α, it can be proved that the estimator for the residual, Δ, is:

$$\widehat{\Delta} = \overline{y} -_H \overline{X^{Bl}}\, \widehat{b_\alpha},$$

or alternatively as

$$\widehat{\Delta} = \overline{y} -_H (\overline{x^M}\, \widehat{\alpha_1} + \overline{x^S}\, \widehat{\alpha_2}).$$

Indeed, as the existence of Hukuhara's difference $y -_H X^{Bl}\, \widehat{b_\alpha}$ is guaranteed, $d^2_\tau(y, X^{Bl}\, \widehat{b_\alpha} + 1\Delta) = d^2_\tau(y -_H X^{Bl}\, \widehat{b_\alpha}, 1\Delta)$ and applying Fréchet principle, it is obtained

$$\widehat{\Delta} = \overline{y -_H (x^M\, \widehat{\alpha_1} + x^S\, \widehat{\alpha_2})} = \overline{y} -_H (\overline{x^M}\, \widehat{\alpha_1} + \overline{x^S}\, \widehat{\alpha_2}).$$

4 Applications: A Comparative Study

The first example is concerned with the relationship between the systolic and diastolic pressures in some patients in the hospital Valle del Nalón, in Asturias. The pulse rate as well as both pressure ranges along a day will be modelled by random intervals, where the endpoints of the interval are the minimum and maximum respectively. The mathematical structure will be given by $\Omega = \{3000 \text{ patients of the hospital}\}$, the Borel σ-algebra and a probability P which is uniformly distributed.

Table 1 represents the data of the sample of 56 patients. For this example the constraint $\operatorname{spr} \boldsymbol{x}\, b_\alpha \leq \operatorname{spr} \boldsymbol{y}$ is fulfilled for the 56 patients. Table 2 summarizes the estimates for α_1 and α_2. For the separate models approach, b_0^m and b_0^s refer to the real-valued intercepts while for the rest of the procedures Δ denotes the interval-valued intercept.

Table 1 $\boldsymbol{y}$: diastolic blood preassure (mmHg) and $\boldsymbol{x}$: systolic blood pressure (mmHg)

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{x}$	$\boldsymbol{y}$
118-173	63-102	119-212	47-93	98-160	47-108
104-161	71-118	122-178	73-105	138-221	70-118
131-186	58-113	127-189	74-125	97-154	60-107
105-157	62-118	113-213	52-112	87-152	50-95
120-179	59-94	141-205	69-133	87-150	47-86
101-194	48-116	99-169	53-109	120-188	53-105
109-174	60-119	126-191	60-98	141-256	77-158
128-210	76-125	99-201	55-121	95-166	54-100
94-145	47-104	88-221	37-94	108-147	62-107
148-201	88-130	94-176	56-121	92-172	45-107
111-192	52-96	102-156	50-94	115-196	65-117
116-201	74-133	103-159	52-95	83-140	45-91
102-167	39-84	102-185	63-118	99-172	42-86
104-161	55-98	111-199	57-113	113-176	57-95
106-167	45-95	130-180	64-121	114-186	46-103
112-162	62-116	103-161	55-97	145-210	100-136
136-201	67-122	125-192	59-101		
90-177	52-104	97-182	54-104		
116-168	58-109	100-161	54-104		
98-157	50-111	159-214	90-127		

All the estimates for α_1 are equal. However, the situation for α_2 is different. Focussing on $\widehat{b_{emp}}$ and $\widehat{b_{exact}}$, it can be seen that they are equal because the sample values fulfil the constraints to assure the existence of the residuals. However, in general, they do not need to be the same value (as shown in next example, due to the fact that $\widehat{b_{exact}}$ was defined to fulfil the constraint, whereas $\widehat{b_{emp}}$ was not). The estimate obtained from the KKT approach is the same as well, but this one was obtained by a numerical approximation. Then we can conclude that the numerical approximation is really close to the exact

Table 2 Estimations of the parameters $\alpha_1, \alpha_2, \Delta$ and b_0^m, b_0^s

Estimator	α_1	α_2	$\Delta/\mathbf{b_0^m - b_0^s}$
$\widehat{\mathbf{b_{exact}}}$	0.4539	0.2570	[1.0164,32.7000]
$\widehat{\mathbf{b_{kkt}}}$	0.4539	0.2570	[1.0164,32.7000]
$\widehat{\mathbf{b_{emp}}}$	0.4539	0.2570	[1.0164,32.7000]
$\widehat{\mathbf{b_{sep}}}$	0.4539	0.6842	16.8582-0.9443

one. $\widehat{b_{sep}}$ reaches a really high value in the estimation of α_2, which seem to denote that this estimator is not a good one, when it is applied to Model M.

The second example is concerned with the relationship between the familiar average income ($\boldsymbol{y}$) and the percentage of people with higher education ($\boldsymbol{x}$) in EEUU in 2006 (http://fact?nder.census.gov). The difference between this example and the previous one is that not all the values of the sample fulfil the constraint $\operatorname{spr} \boldsymbol{x}\, b_\alpha \leq \operatorname{spr} \boldsymbol{y}$. Table 3 displays the data of the sample of 50 people. Then, Table 4 summarizes the values of the different estimates for α_1 and α_2. Again, estimates of α_1 are equal for all the approaches. However, the estimate of α_2 is different for all the approaches excepting the exact and the KKT-based methods.

5 Conclusions

Some approaches to estimate the regression coefficients have been proposed and the comparison between them have been made by means of some examples. According to the empirical results, estimator $\widehat{b_{sep}}$ does not provide good results, which is natural, as they do not account for the specific features of the unified model that has been considered. Thus, $\widehat{b_{sep}}$ will often divert from the $\widehat{b_{exact}}$.

The performance of the empirical estimator depends on the data which has been used. If the data satisfies the constraint to assure the existence of the residuals, then the estimator is similar to the exact one. Otherwise, it is an erroneous estimator, as it provides wrong estimates for α_2, the coefficient accompanying the spreads. In any case, the estimator could be used for large samples, as it approaches to the populational parameter consistenly.
Finally, the numerical estimator $\widehat{b_{kkt}}$ is an adequate, as it reaches values which are really close to the exact ones.

Table 3 $\boldsymbol{y}$: familiar average income, $\boldsymbol{x}$: percentage of people with higher education

State	$\boldsymbol{y}$	$\boldsymbol{x}$	State	$\boldsymbol{y}$	$\boldsymbol{x}$
Alabama	48.460-49.954	7.5-7.9	Alaska	67.501-72.243	8.8-10.2
Arizona	55.063-56.355	8.9-9.5	Arkansas	44.28-45.906	5.9-6.5
California	64.150-64.976	10.3-10.5	Colorado	63.639-65.589	12.1-12.7
Connect.	77.203-79.105	14.0-14.8	Delaware	60.406-64.84	9.9-11.1
Columbia	57.076-65.134	24.4-26.4	Florida	54.043-54.847	8.8-9.0
Georgia	55.503-56.721	9.0-9.4	Hawaii	68.823-71.731	9.3-10.3
Idaho	50.612-52.668	6.8-7.4	Illinois	62.592-63.650	10.6-11.0
Indiana	55.322-56.240	7.8-8.2	Iowa	55.158-56.312	7.1-7.7
Kansas	56.159-57.555	9.5-10.1	Kentucky	48.044-49.408	8.0-8.4
Louisiana	47.467-49.055	6.6-7.0	Maine	51.820-53.766	8.5-9.3
Maryland	76.988-78.690	15.4-16.0	Massach.	73.710-75.216	15.4-15.8
Michigan	57.461-58.531	9.0-9.4	Minnesota	66.324-67.294	9.4-9.8
Mississippi	41.797-43.813	5.8-6.4	Missouri	52.465-53.587	8.5-8.9
Montana	50.177-51.835	7.8-9.0	Nebraska	56.291-57.589	8.0-8.8
Nevada	60.629-62.303	6.9-7.5	N.Hampshire	70.065-72.287	10.6-11.8
N.Jersey	77.226-78.524	12.2-12.6	N.Mexico	46.84749.551	10.5-11.3
N.York	61.774-62.502	13.2-13.4	N.Carolina	51.855-52.817	8.1-8.5
N.Dakota	53.918-56.852	5.9-7.1	Ohio	55.760-56.536	8.1-8.5
Oklahoma	47.179-48.731	7.0-7.4	Oregon	55.166-56.680	9.7-10.3
Pennsylv.	57.787-58.509	9.4-9.8	R.Island	62.762-66.704	10.7-11.9
S.Carolina	49.677-50.991	7.7-8.1	S.Dakota	52.870-54.742	6.7-7.7
Tennes.	49.240-50.368	7.3-7.7	Texas	52.080-52.630	7.9-8.1
Utah	57.306-58.976	9.0-9.8	Vermont	56.752-59.574	12.1-13.5
Virginia	66.263-67.509	12.9-13.5	Washington	63.055-64.355	10.5-10.9
W.Virginia	43.189-44.835	6.3-6.9	Wisconsin	60.172-61.096	8.2-8.6
Wyoming	55.797-59.213	6.8-8.0			

Table 4 Estimates of the parameters

Estimator	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\mathbf{\Delta}}/\mathbf{b}_0^{\hat{\mathbf{m}}} - \mathbf{b}_0^{\hat{\mathbf{s}}}$
$\widehat{\mathbf{b_{exact}}}$	2.9767	1.3817	[29.7003,30.5269]
$\widehat{\mathbf{b_{kkt}}}$	2.9767	1.3817	[29.7003,30.5269]
$\widehat{\mathbf{b_{emp}}}$	2.9767	2.3947	[30.0204,30.2068]
$\widehat{\mathbf{b_{sep}}}$	2.9767	2.6276	30.1136-0.0196

Acknowledgements This research has been partially supported by the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-02 and the Short-Term Scientific Missions associated with the COST Action IC0702 Ref. 010611-008222. Their financial support is gratefully acknowledged. Moreover, part of the work was performed while the first two authors have been visiting the Cyprus University of Technology which partly has supported this research.

References

1. Billard L, Diday E (2003) From the Statistics of data to the Statistics of knowledge: Symbolic Data Analysis. *J Amer Stat Assoc* 98:470–487
2. Blanco-Fernández A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. *Comp Stat Data Anal* 55(9):2568–2578
3. Diamond P (1990) Least squares fitting of compact set-valued data. *J Math Anal Appl* 147:531–544
4. Gil MA, Lubiano A, Montenegro M, López-García MT (2002). Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika* 56:97–111
5. González-Rodríguez G, Blanco A, Corral N, Colubi A (2007) Least squares estimation of linear regression models for convex compact random sets. *Adv D Anal Class* 1:67–81
6. Lima Neto EA, DeCarvalho FAT, Freire ES (2005) Applying constrained linear regression models to predict interval-valued data. *LNAI* 3698:92–106
7. Lima Neto EA, DeCarvalho FAT (2010) Constrained linear regression models for symbolic interval-valued variables. *Comp Stat Data Anal* 54:333–347
8. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Inform Sci* 179(23):3964–3972
9. Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs Reprinted with a detailed "new developments" appendix in 1996 by SIAM Publications, Philadelphia
10. Gil MA, López MT, Lubiano MA, Montenegro M (2001) Regression and correlation analyses of a linear relation between random intervals. *Test* 10(1):183–201 (doi:10.1007/BF02595831)
11. Gil MA, González-Rodríguez G, Colubi A, Montenegro M (2007) Testing linear independence in linear models with interval-valued data. *Computational Statistics and Data Analysis* 51(6):3002–3015. (doi:10.1016/j.csda.2006.01.015)

Hybrid Least-Squares Regression Modelling Using Confidence Bounds

Bülent Tütmez[1] and Uzay Kaymak[2]

Abstract One of the questions regarding bridging of soft computing and statistical methods is the (re-)use of information between the two approaches. In this context, we consider in this paper whether statistical confidence bounds can be used in the hybrid fuzzy least squares regression problem. By using the confidence limits as the spreads of the fuzzy numbers, uncertainty estimates for the fuzzy model can be provided. Experiments have been conducted in the paper, both on regression coefficients and the predicted responses of regression models. The findings show that the use of the confidence intervals as the widths of memberships gives successful results and opens new possibilities in system modeling and analysis.

1 Introduction

Regression is a wide-range statistical methodology used in data analysis. There are many regression methodologies in the literature, such as conventional (ordinary) regression, fuzzy regression and hybrid regression [7, 17]. One of these methods, the hybrid fuzzy regression, is a novel procedure, which combines randomness and fuzziness into a regression model, and can be used in data analysis and system modeling.

The hybrid (fuzzy) regression models proposed in literature are developed based on minimizing fuzziness or the least-squares of errors as the fitting criterion [15, 14]. In addition, some models employ interval analysis [10, 2]. The error term in the regression structures is the main source of the difference between conventional regression and fuzzy regression. In contrast to

[1] School of Engineering, İnönü University, 44280 Malatya, Turkey
bulent.tutmez@inonu.edu.tr
[2] School of Industrial Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands u.kaymak@ieee.org

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 53–63.
springerlink.com

randomness in conventional regression, the error term is considered as fuzzy variables in fuzzy regression. Recently, fuzzy arithmetic based solutions have also been considered [9]. A novel approach of hybrid least-squares regression, which uses the definition of weighted fuzzy arithmetic and the well-accepted least squares fitting criterion, has been proposed by [4].

Both fuzzy set models and statistical models have their specific set of parameters that need to be identified given a data set. For instance, the spreads of fuzzy sets must be determined when applying hybrid regression modeling. An interesting question in this regard is whether information that is defined in the context of statistical models could be useful for developing fuzzy models and vice versa. In this paper, we address one such question, whether the hybrid regression problem can be handled beneficially by using (statistical) confidence intervals. The model uses the hybrid fuzzy least-squares procedure and provides the spreads (widths) of the memberships from confidence bounds. In this manner, integration of statistics and fuzzy analysis is provided in a very specific way. The applicability of confidence interval-based analysis are presented both on linear regression coefficients and predicted values by using numerical examples.

The rest of the paper is organized as follows. Section 2 states the basics of weighted fuzzy arithmetic and the hybrid fuzzy least-squares regression. Confidence interval-based approach for coefficients and predictions is described in Section 3. Section 4 presents the applications of the proposed approach. Finally, Section 5 gives the conclusions.

2 A Review on Hybrid Fuzzy Linear Least-squares Regression

The hybrid fuzzy regression has been developed based on fuzzy arithmetic and least-squares fitting approach. In this section, we give an overview of hybrid regression modeling as formulated in [5, 4].

2.1 Weighted Fuzzy Arithmetic

The approach for hybrid fuzzy linear least-squares regression (LS) is based on the concept of weighted fuzzy arithmetic [4]. Fuzzy arithmetic is a well-known tool used in determining levels of uncertainty [1, 11]. Due to some drawbacks of conventional fuzzy arithmetic, new approaches such as weighted fuzzy arithmetic have been proposed. The weighted fuzzy arithmetic combines each level of operation weighted via the membership level for the entire fuzzy set, and divides the weighted combination using the total integral of the membership function [12]. The method employs the theory of defuzzification

to convert the operation of fuzzy sets into a crisp real number and determines suitable crisp numbers representative (in some sense) of fuzzy sets. The determination of fuzzy widths is the corner stone of the arithmetic operations, since the width controls the fuzziness level of the parameters.

Let $\tilde{A}$ and $\tilde{B}$ be triangular fuzzy numbers which can be denoted as $\tilde{A} = (m_a, c_{a,L}, c_{a,R})$, $\tilde{B} = (m_b, c_{b,L}, c_{b,R})$, respectively, where m is the fuzzy center, c_L is the left fuzzy width, and c_R is the right fuzzy width. At h membership level (i.e. taking the α-cut with $\alpha = h$), the intervals of $\tilde{A}$ and $\tilde{B}$ can be presented as follows [5]:

$$\tilde{A}_h = \left[A_h^L, A_h^R\right] = \left[m_a - (1-h)c_{a,L}, m_a + (1-h)c_{a,R}\right], \tag{1}$$

and

$$\tilde{B}_h = \left[B_h^L, B_h^R\right] = \left[m_b - (1-h)c_{b,L}, m_b + (1-h)c_{b,R}\right]. \tag{2}$$

Based on the intervals above, weighted form of fuzzy addition, fuzzy subtraction, fuzzy multiplication and fuzzy division can be obtained as follows.

$$\tilde{A} + \tilde{B} = (m_a + m_b) + \frac{1}{6}\left[(c_{a,R} + c_{b,R}) - (c_{a,L} + c_{b,L})\right]. \tag{3}$$

$$\tilde{A} - \tilde{B} = (m_a - m_b) + \frac{1}{6}\left[(c_{a,R} - c_{b,R}) - (c_{a,L} - c_{b,L})\right]. \tag{4}$$

$$\begin{aligned}\tilde{A}\tilde{B} = {} & m_a m_b + \left[(m_b c_{a,R} - m_a c_{b,R}) - (m_b c_{a,L} - m_a c_{b,L})\right] + \\ & + \frac{1}{12}(c_{a,L}c_{b,L} + c_{a,R}c_{b,R}).\end{aligned} \tag{5}$$

$$\tilde{A}/\tilde{B} = \int_0^1 \frac{[m_a - (1-h)c_{a,L}]}{[m_b - (1-h)c_{b,L}]}\, h\, dh + \int_0^1 \frac{[m_a + (1-h)c_{a,R}]}{[m_b + (1-h)c_{b,R}]}\, h\, dh. \tag{6}$$

In the system (3) – (6), the coefficients like 1/6 were obtained from the weighted integrations [4].

2.2 Hybrid Multiple Fuzzy Least Squares Regression

Hybrid least-squares system covers fuzzy arithmetic and sum of the squares. By using triangular membership form, the following regression structures can be obtained:

$$\begin{aligned}\hat{Y}_i &= \tilde{A}_0 + \sum_{p'=1}^{p} \tilde{A}_i \tilde{X}_{p',i} \\ &= (a_0, c_{0,L}, c_{0,R}) + (a_1, c_{1,L}, c_{1,R})\, X_{1,i} + \cdots + (a_p, c_{p,L}, c_{p,R})\, X_{p,i},\end{aligned} \tag{7}$$

where p denotes the number of independent variables, $(a_p, c_{p,L}, c_{p,R})$ is the p-th fuzzy slope coefficient, and $(a_0, c_{0,L}, c_{0,R})$ is the fuzzy intercept co-

efficient. In the same way, each measured value $\tilde{Y}_i$ can be presented as $\tilde{Y}_i = (Y_i, e_{i,L}, e_{i,R})$. To formulate the summation of squares of errors between predicted and observed values, minimization of the following E function can be applied [12]:

$$\begin{aligned} E &= \sum_{i=1}^{n} \left(\hat{Y}_i - \tilde{Y}_i \right)^2 \qquad (8) \\ &= \sum_{i=1}^{n} \left\{ (a_0 + a_1 X_i - Y_i)^2 + \frac{1}{3} (a_0 + a_1 X_i - Y_i) \left[(c_{0,R} + c_{1,R} X_i - e_{i,R}) - \right. \right. \\ &\quad \left. - (c_{0,L} + c_{1,L} X_i - e_{i,L}) \right] + \\ &\quad \left. + \frac{1}{12} \left[(c_{0,L} + c_{1,L} X_i - e_{i,L})^2 + (c_{0,R} + c_{1,R} X_i - e_{i,R})^2 \right] \right\}. \end{aligned}$$

The solution procedure, which is similar to the solution of ordinary regression problem, consists of three components: normal equations for fuzzy centers, normal equations for left fuzzy spreads, and normal equations for right fuzzy spreads. For a given set of data $(X_{1,i}, X_{2,i}, \ldots, X_{p,i} : (Y_i, e_{i,L}, e_{i,R}), i = 1, \ldots, n)$, the following normal equations for fuzzy centers can be constructed [4]:

$$\begin{aligned} n\, a_0 + \sum_{i=1}^{n} X_{1,i} a_1 + \sum_{i=1}^{n} X_{2,i} a_2 + \cdots + \sum_{i=1}^{n} X_{p,i} a_p &= \sum_{i=1}^{n} Y_i, \\ \sum_{i=1}^{n} X_{1,i} a_0 + \sum_{i=1}^{n} X_{1,i}^2 a_1 + \sum_{i=1}^{n} X_{1,i} X_{2,i} a_2 + \cdots + \sum_{i=1}^{n} X_{1,i} X_{p,i} a_p &= \sum_{i=1}^{n} X_{1,i} Y_i, \qquad (9) \\ \vdots \quad + \quad \vdots \quad + \quad \vdots \quad + \quad \cdots \quad + \quad \vdots \quad & \quad \vdots \\ \sum_{i=1}^{n} X_{p,i} a_0 + \sum_{i=1}^{n} X_{p,i} X_{1,i} a_1 + \sum_{i=1}^{n} X_{p,i} X_{2,i} a_2 + \cdots + \sum_{i=1}^{n} X_{p,i}^2 a_p &= \sum_{i=1}^{n} X_{p,i} Y_i. \end{aligned}$$

The normal equations for the left spread can be presented as follows:

$$\begin{aligned} n\, c_{0,L} + \sum_{i=1}^{n} X_{1,i} c_{1,L} + \sum_{i=1}^{n} X_{2,i} c_{2,L} + \cdots + \sum_{i=1}^{n} X_{p,i} c_{p,L} &= \sum_{i=1}^{n} e_{i,L}, \qquad (10) \\ \sum_{i=1}^{n} X_{1,i} c_{0,L} + \sum_{i=1}^{n} X_{1,i}^2 c_{1,L} + \sum_{i=1}^{n} X_{1,i} X_{2,i} c_{2,L} + \cdots + \sum_{i=1}^{n} X_{1,i} X_{p,i} c_{p,L} &= \sum_{i=1}^{n} X_{1,i} e_{i,L}, \\ \vdots \quad + \quad \vdots \quad + \quad \vdots \quad + \quad \cdots \quad + \quad \vdots \quad & \quad \vdots \\ \sum_{i=1}^{n} X_{p,i} c_{0,L} + \sum_{i=1}^{n} X_{p,i} X_{1,i} c_{1,L} + \sum_{i=1}^{n} X_{p,i} X_{2,i} c_{2,L} + \cdots + \sum_{i=1}^{n} X_{p,i}^2 c_{p,L} &= \sum_{i=1}^{n} X_{p,i} e_{i,L}. \end{aligned}$$

The normal equations for the right spread can be presented as follows:

$$n\,c_{0,R} + \sum_{i=1}^{n} X_{1,i}c_{1,R} + \sum_{i=1}^{n} X_{2,i}c_{2,R} + \cdots + \sum_{i=1}^{n} X_{p,i}c_{p,R} = \sum_{i=1}^{n} e_{i,R}, \quad (11)$$

$$\sum_{i=1}^{n} X_{1,i}c_{0,R} + \sum_{i=1}^{n} X_{1,i}^2 c_{1,R} + \sum_{i=1}^{n} X_{1,i}X_{2,i}c_{2,R} + \cdots + \sum_{i=1}^{n} X_{1,i}X_{p,i}c_{p,R} = \sum_{i=1}^{n} X_{1,i}e_{i,R},$$

$$\vdots \quad + \quad \vdots \quad + \quad \vdots \quad + \quad \cdots \quad + \quad \vdots \qquad \vdots$$

$$\sum_{i=1}^{n} X_{p,i}c_{0,R} + \sum_{i=1}^{n} X_{p,i}X_{1,i}c_{1,R} + \sum_{i=1}^{n} X_{p,i}X_{2,i}c_{2,R} + \cdots + \sum_{i=1}^{n} X_{p,i}^2 c_{p,R} = \sum_{i=1}^{n} X_{p,i}e_{i,R}.$$

3 Confidence Interval-based Approach

In the procedures given in (9) – (11), the fuzzy centers of fuzzy regression coefficients are provided by the fuzzy center of observed data. The left fuzzy spreads and the right fuzzy spreads of the fuzzy regression coefficients are obtained separately from the corresponding fuzzy spreads of the observed data [5].

In the above system, determining fuzzy widths has crucial importance on the solution. In this paper, we investigate the use of confidence intervals as a tool to determine the spreads of fuzzy sets. For the hybrid regression system, two types of confidence intervals can be considered.

- Confidence intervals on the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the general multilinear model.
- Prediction interval on $Y|x_1, x_2, \ldots, x_p$, an individual response for a given set of values of the predictor variables.

3.1 Confidence Interval on Coefficients

As discussed in [6], to provide the regression coefficients, the fuzzy regression methods can use various parameters such as maximum compatibility criterion, minimum fuzziness criterion, or interval analysis. Similarly, the use of confidence-bounds for computing the coefficients in hybrid linear system could be an alternative way to be able to extend a point estimate for a parameter to an interval estimate in the way of appraisal of its accuracy [3].

In a general linear model, the estimators $\beta_0, \beta_1, \ldots, \beta_p$ can be taken as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \quad (12)$$

The variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. The variances of β_0, β_1, …, β_p can be stated by $c_{00}\sigma^2, c_{11}\sigma^2, c_{pp}\sigma^2$, where c_{ii} represents the element on the main diagonal in row $i+1$ of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. From this point,

the standard normal form of the random variable can be presented as follows:

$$Z = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}}, \tag{13}$$

where σ is unknown. We can replace it with its estimator $S = \sqrt{SSE/(n-p-1)}$ to form the T_{n-p-1} random variable

$$T_{n-p-1} = \frac{\hat{\beta}_i - \beta_i}{S\sqrt{c_{ii}}}. \tag{14}$$

The residual or error sum of squares, SSE, may be large either because Y exhibits a high variability naturally or because the assumed model is inappropriate. Based on the equations above, the confidence bound for β_i, which is the ith model parameter in the general linear model, can be expressed as follows [13]:

$$\hat{\beta}_i \pm t_{\alpha/2} S\sqrt{c_{ii}}, \tag{15}$$

where the point $t_{\alpha/2}$ is the appropriate point based on the T_{n-p-1} distribution.

3.2 Prediction Interval on a Single Predicted Response

From hybrid regression perspective, one of the most useful types of confidence intervals is those on single predicted responses. The confidence bounds for an individual response for a given set of values of predictor variables can be considered in the hybrid-least squares problem. From this point of view, the prediction bounds for $\hat{Y}|x_{10}, x_{20}, \ldots, x_{p0}$, an individual response for a given set of values of the predictor variables, can be stated as follows [13]:

$$\hat{Y}|x_{10}, x_{20}, \ldots, x_{p0} \pm t_{\alpha/2} S\sqrt{1 + \mathbf{x_0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}, \tag{16}$$

where the point $t_{\alpha/2}$ is the appropriate point based on the T_{n-p-1} distribution. The expression given in (16) can provide the spreads of the fuzzy numbers like triangular numbers, directly. Therefore, the error values employed in the hybrid least squares system can be obtained using the prediction intervals on a different confidence levels such as 90% or 95%, depending on the user preference and problem type.

4 Experimental Studies

The experimental studies were conducted for both regression coefficients and prediction values. The studies performed by real data sets derived from geo-environmental science and reliability measures of the developed models have been conducted.

4.1 Experiments on Regression Coefficients

The methodology proposed for evaluating the regression coefficients is close to the fuzzy least-squares approach by minimum fuzziness criterion. Two stages were applied in computing regression coefficients. In the first stage, the fuzzy centers have been provided from the ordinary least-squares regression. After that, by using (15), the widths of memberships (coefficients) have been obtained.

The application has been shown using a basic regression data [10]. This data set was preferred, because some experiments were carried out on this data in the literature, so that we can conduct a comparison. In the following, data for symmetrical triangular fuzzy numbers are considered. It covers the following crisp X and crisp Y values:

$$[(X_i : Y_i)] = [(2 : 14), (4 : 16), (6 : 14), (8 : 18), \\ (10 : 18), (12 : 22), (14 : 18), (16 : 22)].$$

By the fuzziness using confidence bounds, the regression expressions have been provided at different levels of confidence which are 80%, 90% and 95%. Table 1 gives the regression equations together with the results from former works that were presented in [6]. The equations consist of the resulting mean and width values. It can be seen that the solutions from hybrid LS regression (first number in the brackets) are centered around the ordinary LS regression solutions, but the uncertainty associated with the solution (second number in the brackets) has been reduced compared to solutions from fuzzy LS regression with the minimum fuzziness criterion. Furthermore, the associated fuzziness increases with increasing confidence levels, as expected. Therefore, more uncertainty must be accepted in order to have larger confidence.

4.2 Experiments on Predicted Values

To illustrate the prediction bounds on predicted observations, two experiments have been conducted by real data sets. The performance assessments of the models have been carried out using reliability measures. The hybrid

Table 1 Regression coefficients obtained from different models.

Regression method	Hybrid regression equation
Ordinary LS regression	$\hat{Y} = 12.93 + 0.54X$
Fuzzy LS regression using maximum compatibility criterion	$\hat{Y} = 12.55 + 0.59X \pm \sqrt{229.86}X + 3.08X^2$
Fuzzy LS regression using minimum fuzziness criterion	$\hat{Y} = (12.93, 5.83) + (0.54, 0.23)X$
Hybrid LS regression using 95% confidence-bounds	$\hat{Y} = (12.93, 3.43) + (0.54, 0.34)X$
Hybrid LS regression using 90% confidence-bounds	$\hat{Y} = (12.93, 2.73) + (0.54, 0.27)X$
Hybrid LS regression using 80% confidence-bounds	$\hat{Y} = (12.93, 2.02) + (0.54, 0.20)X$

correlation coefficient (HR), which shows the linearity assumption of the hybrid model, is employed for reliability evaluation. In addition, a hybrid standard error of estimates (HS_e) is employed to measure the goodness of fit between the hybrid regression model and measured data. In particular, the smaller HS_e indicates the better goodness of fit and better accuracy of predictions. Authors of [5] formulated the HR and HS_e using the weighted fuzzy arithmetic as follows:

$$HR^2 = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(\tilde{Y}_i - \bar{Y}\right)^2}, \tag{17}$$

$$HS_e = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(\hat{Y}_i - \tilde{Y}_i\right)^2}, \tag{18}$$

in which $n-p-1$ is the degrees of freedom. $\tilde{Y}$ and $\hat{Y}$ represent measured and predicted values, respectively. $\bar{Y}$ denotes the mean of the measured values. Note that the open forms of the expressions above include the fuzzy terms [4].

4.2.1 Application 1

In the first experimental work, Oasis Valley data set has been used. The U.S. government has exploded a large number of nuclear devices underground at the Nevada Test Site, many at or below the water table. The data set provides a table for 19 trace elements in water collected from 22 wells and springs in the Nevada Test Site and adjacent Oasis Valley [8].

The data covers the locations of the wells and springs in arbitrary Cartesian coordinates. Because the relationships between the coordinates and trace

elements are important for geo-environmental appraisal, the locations have been considered as predictors and trace elements Lithium (Li), Rubidium (Rb) and Selenium (Se) have been determined as response variables for three different experiments, respectively. The applications have been conducted at 95% level of confidence. To perform the reliability analysis, both conventional reliability measures and hybrid reliability measures were computed. Both the results are summarized in Table 2 for each trace element.

Table 2 Reliability measures for trace element models.

Trace element	Conventional reliability measure R^2	S_e	Hybrid reliability measure HR^2	HS_e
Li	0.500	38.886	0.547	38.529
Rb	0.477	4.590	0.667	4.535
Se	0.113	0.324	0.485	0.316

4.2.2 Application 2

In the second application, a coal site has been considered. Sivas-Kalburçayırı field is one of the most important lignite reserves in Turkey. Lignite seams in this field are utilized to feed coal to a power plant. The data set comprises of 67 records of the field [16]. Lignite quality (calorific value) is mainly controlled by four factors which are moisture content, ash content, volatile matter or sulphur content. Therefore a multi-linear structure has been designed using these variables. The application has been performed at the 90% level of confidence. As a result of the calculations, the parameters for reliability measure have been obtained as in Table 3.

Table 3 Reliability measures for lignite quality.

Response	Conventional reliability measure R^2	S_e	Hybrid reliability measure HR^2	HS_e
Calorific value	0.809	128.500	0.835	129.455

4.2.3 Discussion

The experiments on the confidence interval-based model suggest that the procedure obtains a convenient tool for dealing with crisp measured data by analyzing the random errors based on a widely accepted statistical ground. In

addition, the method enables an opportunity to analyze a system on different levels of confidence and also provides a transparent way for uncertainty evaluation.

If HS_e is greater than the standard deviation of measured values of response variable $S_{\tilde{y}}$, the regression analysis can not be accepted as successful. Because the $S_{\tilde{y}}$ is a constant and is independent from the method, the ratio $HS_e/S_{\tilde{y}}$ is a normalized measure of the goodness of fit [4]. Thus, HS_e, $HS_e/S_{\tilde{y}}$ and HR measures can be employed in performance evaluations.

As can be seen in Tables 2 and 3, HR^2 values of hybrid model is better than conventional regression for both models. This finding can be confirmed for the second case study using HS_e. The confidence-based analysis leads to improved HS_e values in the first case study. $HS_e/S_{\tilde{y}}$ ratios have been obtained as 0.888, 0.791 and 1.0 for ordinary regression of each trace element, respectively. The values for confidence-bound based approach have been recorded as 0.880, 0.782 and 0.975, respectively. A small ratio is preferred in modelling works, and most of the applications satisfy this condition. In the second study, the ratio has been obtained as 0.951 and 0.956 for ordinary regression and hybrid regression, respectively. Due to the close values recorded in the application, no meaningful difference between the methods based on $HS_e/S_{\tilde{y}}$ can be mentioned.

Two drawbacks of the study should be mentioned. First, the experiments have been conducted using crisp input and outputs. Therefore, the performances on the fuzzy inputs and outputs have not been evaluated. Second, the data was limited and very heterogeneous (natural field data), because of which there was no consistent performance improvement owing to the model. However, the approach given in this study and the former works in literature may provide some possibilities for future investigations on large and conditioned data sets.

5 Conclusions

Hybrid fuzzy least squares regression problem was handled using confidence bound, which is a well-known statistical procedure. Instead of the errors provided by some formulations, the error based confidence-intervals have been suggested in hybrid least squares regression formulation. The experimental works with real data sets have been performed, both with regression coefficients and with predicted responses. The results showed that the confidence interval-based approach can provide successful results and also possibilities for future works in depth.

References

1. Bardossy G, Fodor J (2002) *Evaluation of Uncertainties and Risks in Geology.* Springer, Berlin
2. Blanco-Fernández A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics & Data Analysis* 55(9):2568–2578
3. Bowermann BL, O'Connell R (2000) *Linear Statistical Models: an applied approach.* Duxbury Press, Boston, USA
4. Chang YHO (2001) Hybrid fuzzy least-squares regression analysis and its reliability measures. *Fuzzy Sets and Systems* 119:225–246
5. Chang YHO, Ayyub BM (1998) Hybrid least-squares regression analysis. In: Ayyub BM, Gupta MM (eds.) *Uncertainty Analysis in Engineering and Sciences, International Series in Intelligent Technologies, vol 11*, chap 12. Kluwer Academic Publishers, Boston
6. Chang YHO, Ayyub BM (2001) Fuzzy regression methods — a comparative assessment. *Fuzzy Sets and Systems* 119:187–203
7. Draper N, Smith H (1998) *Applied Regression Analysis.* Wiley, New York
8. Farnham IM, Stetzenbach KJ, Singh AK, Johannesson KH (2000) Deciphering groundwater flow systems in Oasis Valley, Nevada, using trace element chemistry, multivariate statistics, and geographical system. *Mathematical Geology* 32(8):943–968
9. González-Rodríguez G, Blanco A, Colubi A, Lubiano A (2009) Estimation of a simple linear regression model for fuzzy random variables. *Fuzzy Sets and Systems* 160(3):357–370
10. Ishibuchi H (1992) Fuzzy regression analysis. *Japan Journal of Fuzzy Theory and Systems* 4:137–148
11. Kaufmann A, Gupta MM (1991) *Introduction to Fuzzy Arithmetic — theory and applications.* Van Nostrand Reinhold Company, New York
12. Kwong CK, Chen Y, Chan KY, Wong H (2008) The hybrid fuzzy least-squares regression approach to modeling manufacturing processes. *IEEE Trans. on Fuzzy Systems* 16(3):644–651
13. Milton JS, Arnold JC (1995) *Introduction to Probability and Statistics.* McGraw-Hill, Singapore
14. Savic D, Pedrycz W (1991) Evaluation of fuzzy regression models. *Fuzzy Sets and Systems* 39:51–63
15. Tanaka H, Uejima S, Asai K (1982) Linear regression analysis with fuzzy model. *IEEE Trans. on Systems, Man and Cybernetics* 12(6):903–907
16. Tercan AE, Karayiğit AI (2001) Estimation of lignite reserve in the Kalburcayiri field, Kangal basin, Sivas, Turkey. *Int. Journal of Coal Geology* 47:91–100
17. Yang MS, Liu HH (2003) Fuzzy least-squares algorithms for interactive fuzzy linear regression models. *Fuzzy Sets and Systems* 135:305–316

Testing the Variability of Interval Data: An Application to Tidal Fluctuation

Ana Belén Ramos-Guajardo[1] and Gil González-Rodríguez[1]

Abstract A methodology for analyzing the variability of the tidal fluctuation in a specific area is proposed in this work. The idea is to consider intervals determined by the minimum and maximum height reached by the tides in a day. Thus, the theoretical aim is to develop hypothesis tests about the variance of one or more interval-valued random elements (i.e., *random intervals*). Some simulations showing the empirical behavior and consistency of the proposed tests are carried out by considering different models. Finally, the procedure is applied to a real-life study concerning the fluctuation of tides in the port of Gijón (Asturias).

1 Introduction

Tides refer to the vertical motion of water caused by the gravitational effects of the sun and moon. They vary on timescales ranging from hours to years due to numerous influences, called tidal constituents, such as rotation of the Earth, the positions of the moon and the sun relative to Earth, moon altitude (elevation) above the Earth equator, bathymetry, etc. From ancient times, tidal observation and discussion has increased in sophistication, first marking the daily recurrence, then the relationship with the sun and moon. Some studies concerning the tidal fluctuation have been developed in the literature along the last years (see, for instance, [4, 8, 21]).

One of the most important applications of tidal fluctuation is tidal power, which is a form of hydropower that converts the energy of tides into useful forms of power (mainly electricity). Thus, the energy of tidal flows is converted into electricity by using a tidal generator. Some studies about this topic can be found in [3, 9, 7]. In this context, it is interesting to analyze, for

[1] Dpto. de Estadística, I.O. y D.M., Universidad de Oviedo, 33007 Oviedo, Spain
ramosana@uniovi.es · gil@uniovi.es

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 65–74.
springerlink.com

instance, if there has been a big variation of the tidal fluctuation in a specific year with respect to the previous year, in order to determine the quantity of electricity that can be generated by the energy of tidal flows.

It is well-known that two high tides and two low tides are usually observed each day. Thus, experimental data regarding tidal fluctuations is usually given by the heights reached in the two high tides as well as the heights reached in the two low tides. To sum up these observations in an unique value would entail a loss of information that can be avoided by using compact intervals determined by the minimum and the maximum heights that the tides reach in a day at a specific place. Some studies regarding interval data can be found, for instance, in [6, 11, 22].

Random intervals have been shown to be useful in handling this kind of random elements (see, for instance, [17, 12, 15]), when the interest is focussed on the associated random element for which values are intervals instead of on the real-valued random variables determining either the extremes of the intervals or inner points within them. They were introduced to formalize imprecise experimental data which can be described by means of intervals. Thus, random intervals associate compact intervals with experimental outcomes.

The aim of this paper is to test whether or not the variance of a random interval is equal to, greater than or lower than a given value. One-sample tests for a Fréchet-type variance in the fuzzy context have been previously developed by taking into account different situations in Lubiano *et al.* [18] and in Ramos-Guajardo *et al.* [23]. Here, the idea is to derive from the these studies the corresponding results for the interval case by considering a generalized metric introduced in [24], in order to apply them to the proposed real-life situation.

Some preliminaries are first recalled in Section 2. A brief review of statistical hypothesis tests about the variance of a random interval is then presented in Section 3. In Section 4 some simulations of the proposed tests are reported. Section 5 shows an application to a real life study about the variability of tidal fluctuations in the port of Gijón (Asturias). Finally, some conclusions and current lines of research are gathered in Section 6.

2 Preliminaries

Let $\mathcal{K}_c(\mathbb{R})$ be the class of the nonempty convex compact subsets of $\mathbb{R}$. An interval $A \in \mathcal{K}_c(\mathbb{R})$ can be expressed in two different ways. The first one is based on the infimum and the supremum of the interval ($A = [\inf A, \sup A]$) satisfying the order constraint $\inf A \leq \sup A$. Due to the difficulties arising with order restrictions, it is often more convenient to work with the *mid/spread parametrization of* A, $(\mathrm{mid}\, , \mathrm{spr}\,) \in \mathbb{R} \times \mathbb{R}^+$, which is defined as

$$\mathrm{mid}\, A = (\sup A + \inf A)/2 \quad \text{and} \quad \mathrm{spr}\, A = (\sup A - \inf A)/2.$$

The class $\mathcal{K}_c(\mathbb{R})$ can be naturally endowed with an inner composition law and an external one which are, respectively, the Minkowski addition and the product by a scalar, so that for all intervals $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$,

- $A+B = [\text{mid}\, A+\text{mid}\, B-(\text{spr}\, A+\text{spr}\, B), \text{mid}\, A+\text{mid}\, B-(\text{spr}\, A+\text{spr}\, B)]$,
- $\lambda \cdot A = [\lambda \text{mid}\, A - |\lambda| \text{spr}\, A, \lambda \text{mid}\, A + |\lambda| \,\text{spr}\, A]$.

The space $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear due to the lack of a symmetric element with respect to the Minkowski addition. To overcome this problem it is useful to consider the so-called *Hukuhara difference* $A -_H B$, which is defined as the set difference C, provided that $C \in \mathcal{K}_c(\mathbb{R})$, so that $A = B + C$. Nevertheless, such a difference is not generally well-defined; in fact, given $A, B \in \mathcal{K}_c(\mathbb{R})$, it is is easy to show that $A -_H B$ exists if and only if $\text{spr}\, B \leq \text{spr}\, A$.

Formally, if $(\Omega, \mathcal{A}, P)$ is a probability space, a *random interval* can be defined as a Borel measurable mapping $X : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ with respect to the σ-field generated by the topology induced by the well-known Hausdorff metric d_H on $\mathcal{K}_c(\mathbb{R})$. Equivalently, X is a random interval if and only if $\text{mid}\, X$ and $\text{spr}\, X$ are real-valued random variables.

Now some summarizing measures concerning random intervals are introduced. The first one concerns a central tendency measure and it is the expected value of a random interval (which is also an interval). Thus, if X is a random interval verifying that $E(|X|) < \infty$ (with $|X|(w) = \sup\{|x| \text{ s.t. } x \in X(w) \text{ for } \omega \in \Omega\}$), then the expected value of X in Kudo-Aumann's sense (see, e.g., [2]) is given by

$$E(X) = \left\{ E(f) \,|\, f : \Omega \to \mathbb{R}\,,\; f \in L^1((\Omega, \mathcal{A}, P))\,,\; f \in X \;a.s.[P] \right\}.$$

Equivalently, $E(X)$ can be expressed as

$$E(X) = [E(\text{mid}\, X) - E(\text{spr}\, X), E(\text{mid}\, X) + E(\text{spr}\, X)].$$

Some good properties of the expected value of a random interval are that it is linear and it is coherent with the arithmetic considered for finite populations and in the sense of the Strong Law of Large Numbers (see [1]).

The second one concerns the dispersion of the random intervals, which is measured by means of a real-valued *variance* quantified in terms of mean 'error'. To define this variance we have considered before a generalized distance which was firstly proposed in [12] (see also [24]), and such that, for $A, B \in \mathcal{K}_c(\mathbb{R})$, its square has the following expression:

$$d_\theta^2(A, B) = (\text{mid}\, A - \text{mid}\, B)^2 + \theta(\text{spr}\, A - \text{spr}\, B)^2,$$

where $\theta > 0$ determines the relative weight of the distance between the spreads against the distance between the mids. Thus, the choice of θ will assign less, equal or more weight to the distances between the spreads (connected with the imprecision of the intervals) than to the distances between the mids (connected with the location of the intervals). Then, provided that

$E(|X|^2) < \infty$, the *variance* of a random interval X is defined in a Fréchet sense (see [10, 16]) as

$$\sigma_X^2 = E\left(d_\theta^2(X, E(X))\right).$$

Finally, the corresponding sample moments are defined as in the classical case. Consider a simple random sample of size n from the random interval X, $\{X_1, \ldots, X_n\}$ (i.e., $X_1, \ldots, X_n$ are independent random intervals and identically distributed as X). Then,

- The *interval-valued sample mean of* $\{X_i\}_{i=1}^n$ is given by $\overline{X_n} = (X_1 + \ldots + X_n)/n$.
- The *real-valued Fréchet sample variance of* $\{X_i\}_{i=1}^n$ is given by $\widehat{\sigma}_X^2 = \sum_{i=1}^n d_\theta^2(X_i, \overline{X_n})/n$, although considering the *real-valued sample quasi-variance of* $\{X_i\}_{i=1}^n$, $\widehat{S}_X^2 = \sum_{i=1}^n d_\theta^2(X_i, \overline{X_n})/(n-1)$, would be even preferable because it is an unbiased and consistent estimator of the population variance (see [20]).

3 One-sample Tests for the Variance of a Random Interval

Let $X_1, \ldots, X_n$ be n independent and identically distributed random intervals. Given $\sigma_0 \in \mathbb{R}^+$, the aim of this work is to solve the following tests:

$$H_0^A : \sigma_X^2 = \sigma_0^2 \; vs. \; H_1^A : \sigma_X^2 \neq \sigma_0^2 \qquad \text{(Problem 1)}; \tag{1}$$

$$H_0^B : \sigma_X^2 \leq \sigma_0^2 \; vs. \; H_1^B : \sigma_X^2 > \sigma_0^2 \qquad \text{(Problem 2)}; \tag{2}$$

$$H_0^C : \sigma_X^2 \geq \sigma_0^2 \; vs. \; H_1^C : \sigma_X^2 < \sigma_0^2 \qquad \text{(Problem 3)}. \tag{3}$$

To design tests for these problems, the following statistic can be considered by taking into account the results provided in [23]:

$$T_n = \frac{\sqrt{n}\left(\widehat{S}_X^2 - \sigma_0^2\right)}{\sqrt{\widehat{\sigma}^2_{d_\theta^2(X, \overline{X_n})}}}. \tag{4}$$

It is straightforward to prove that $\widehat{S}_X^2 = \widehat{S}^2_{\text{mid}\,X} + \theta \widehat{S}^2_{\text{spr}\,X}$ and $\sigma_X^2 = \sigma^2_{\text{mid}\,X} + \theta \sigma^2_{\text{spr}\,X}$. Thus, the Central Limit Theorem for real-valued variables shows that

$$T_n^1 = \sqrt{n}\left(\widehat{S}_X^2 - \sigma_0^2\right) = \sqrt{n}\left(\left(\widehat{S}^2_{\text{mid}\,X} - \sigma^2_{\text{mid}\,X}\right) + \theta\left(\widehat{S}^2_{\text{spr}\,X} - \sigma^2_{\text{spr}\,X}\right)\right)$$

converges in law to a $\mathcal{N}\left(0, \sigma_{d_\theta^2(X,E(X))}\right)$ whenever the condition $E(|X|^4) < \infty$ is fulfilled.

In addition, it is easy to show that $\widehat{\sigma}^2_{d_\theta^2(X,\overline{X_n})} \xrightarrow{c.s.} \sigma^2_{d_\theta^2(X,E(X))}$ by considering Proposition 3 provided in [23]. Therefore,

$$T_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Given $\alpha \in [0,1]$, the following asymptotic testing procedure is provided.

- Let $z_{1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the distribution $\mathcal{N}(0,1)$. If H_0 in Problem 1 is true, then $\lim_{n\to\infty} P\left(|T_n| > z_{1-\alpha/2}\right) = \alpha$. Thus, the test that consist in rejecting H_0 in Problem 1 when $T_n > z_{1-\alpha/2}$ is asymptotically correct.
- Let $z_{1-\alpha}$ be the $(1-\alpha)$-quantile of the distribution $\mathcal{N}(0,1)$. If H_0 in Problem 2 is true, then $\limsup_{n\to\infty} P\left(T_n > z_{1-\alpha}\right) \leq \alpha$, and the equality is reached for $\sigma^2_{\mathcal{X}} = \sigma^2_0$. Thus, the test that consist in rejecting H_0 in Problem 2 when $T_n > z_{1-\alpha}$ is asymptotically correct.
- Let z_α be the α-quantile of the distribution $\mathcal{N}(0,1)$. If H_0 in Problem 3 is true, then $\limsup_{n\to\infty} P\left(T_n < z_\alpha\right) \leq \alpha$, and the equality is reached for $\sigma^2_{\mathcal{X}} = \sigma^2_0$. Thus, the test that consist in rejecting H_0 in Problem 3 when $T_n < z_\alpha$ is asymptotically correct.

On the other hand, the application of bootstrap techniques to solve hypothesis tests has been shown to provide better results than the asymptotic ones when imprecise data are involved (see, for instance, [19, 13, 14]).

Suppose that X is a random interval and that $\{X_i\}_{i=1}^n$ is a simple random sample drawn from X. Let $\{X_i^*\}_{i=1}^n$ be a bootstrap sample from $\{X_i\}_{i=1}^n$. Then, by considering analogous developments than the ones in [23], the following bootstrap statistic can be considered:

$$T_n^* = \frac{\sqrt{n}\left(\widehat{S}^2_{X^*} - \widehat{S}^2_X\right)}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(d_\theta^2\left(X_i^*, \overline{X_n^*}\right) - \widehat{\sigma}^2_{X^*}\right)^2}}. \tag{5}$$

The previous statistic is an approximation of the distribution of T_n in the worst situation under H_0. Finally, the Monte Carlo method is used in order to approximate the unknown distribution of T_n^*.

4 Simulation Studies

To show the performance of the proposed tests, some simulations are carried out. Specifically, Problems 1, 2 and 3 are analyzed. Given two random

intervals X and Y, two different situations are considered depending on the distributions chosen for the mid and the spread of X and Y, namely,

- **Case1:** $\operatorname{mid} X \equiv U(0,15)$ and $\operatorname{spr} X \equiv U(0,15)$.
- **Case2:** $\operatorname{mid} Y \equiv U(0,10)$ and $\operatorname{spr} Y \equiv \operatorname{mid} Y + \beta(1,5)$.

It is easy to show that $\sigma_X^2 = 25$ and $\sigma_Y^2 = 11.18$, so that $\sigma_0^2 = 25$ in Case 1 and $\sigma_0^2 = 11.18$ in Case 2. Simple random samples of different sizes were drawn from X and Y, respectively. In all the cases, the value chosen for θ was $1/3$ which is the weight associated with the Lebesgue measure λ on $[0,1]$ when the equivalence of Bertoluzza metric and d_θ is considered (see [24]).

Firstly, the results for the asymptotic case are presented. 10,000 simulations of the asymptotic testing procedures have been carried out and the results for different sample sizes n and the usual significance levels are reported in Tables 1 and 2.

Table 1: Empirical percentage of rejections under H_0 (asymptotic tests, Case 1)

	$H_0 : \sigma_X^2 = 25$			$H_0 : \sigma_X^2 \geq 25$			$H_0 : \sigma_X^2 \leq 25$		
$n \setminus 100\,\beta$	1	5	10	1	5	10	1	5	10
50	1.55	6.05	11.92	2.44	9.05	15.11	.47	2.66	6.73
100	1.33	5.76	10.49	1.81	7.78	13.27	.57	3.88	7.68
500	.97	5.22	9.96	1.5	5.69	11.05	.71	3.97	8.7
1,000	1.07	5.15	10.34	1.01	5.66	10.46	.89	4.41	9.32
5,000	1.09	4.83	9.88	1.11	5.19	9.96	.98	4.92	9.86

Table 2: Empirical percentage of rejections under H_0 (asymptotic tests, Case 2)

	$H_0 : \sigma_Y^2 = 11.18$			$H_0 : \sigma_Y^2 \geq 11.18$			$H_0 : \sigma_Y^2 \leq 11.18$		
$n \setminus 100\,\beta$	1	5	10	1	5	10	1	5	10
50	2.03	6.69	11.69	3.17	8.6	15	.58	3.16	6.78
100	1.3	5.56	10.8	1.96	7.25	13.25	.73	3.78	7.62
500	.99	5.1	10.08	1.3	5.19	11.13	.89	4.54	9.49
1,000	1.12	4.97	10.17	1.16	5.16	10.64	.91	4.87	9.72
5,000	1.03	5.05	9.95	1.06	5.15	10.17	.98	4.95	9.92

On the other hand, 10,000 simulations of the bootstrap tests have been performed at the usual significance levels and different sample sizes n, with 1,000 bootstrap replications. The results are shown in Tables 3 and 4.

Tables 1 and 2 show that the asymptotic procedure requires large sample sizes, since only when $n \geq 1000$ the empirical percentage of rejections is

Table 3: Empirical percentage of rejections under H_0 (bootstrap tests, Case 1)

	$H_0 : \sigma^2_Y = 11.18$			$H_0 : \sigma^2_Y \geq 11.18$			$H_0 : \sigma^2_Y \leq 11.18$		
$n \setminus 100\,\beta$	1	5	10	1	5	10	1	5	10
10	.66	3.54	7.4	1.84	10.64	18.72	2.16	9.66	19.18
30	.92	4.12	8.1	1.6	7.34	14.06	1.86	7.58	14.6
50	.93	4.8	9.64	1.26	6.58	12.6	1.52	6.58	12.82
100	1.02	5.04	9.92	1.18	5.84	11.5	1.4	5.78	11.08
200	1.06	4.92	9.98	1.09	5.42	10.8	1.14	5.32	10.5

Table 4: Empirical percentage of rejections under H_0 (bootstrap tests, Case 2)

	$H_0 : \sigma^2_Y = 11.18$			$H_0 : \sigma^2_Y \geq 11.18$			$H_0 : \sigma^2_Y \leq 11.18$		
$n \setminus 100\,\beta$	1	5	10	1	5	10	1	5	10
10	.62	3.72	7.74	1.22	8.78	15.92	1.16	8.48	16.58
30	.86	4.02	9.06	1.58	7.01	13.8	1.46	7.08	13.14
50	.92	4.88	9.62	1.2	6.12	12.02	1.28	6.02	11.65
100	.94	4.92	9.78	1.12	5.6	10.84	1.13	5.65	10.35
200	.1	4.96	9.96	1.08	5.28	10.42	1.03	5.25	10.21

quite close to the nominal significance level. Nevertheless, the results of both bootstrap tests are quite good from $n \geq 100$ as shown in Tables 3 and 4, since the empirical percentage of rejections is quite close to the nominal significance level in both cases from a sample size of 100. In addition, in the two-sample case the bootstrap approach behaves "quite good" when a sample size of 50 individuals is considered.

5 Application to Tidal Fluctuations

To illustrate in practice the procedure proposed in previous sections, its application is shown in a real-life situation regarding the tidal fluctuations obtained along 2010 in the port of Gijón (Asturias, Spain). Data can be found in `<http://lapescasubmarina.com/universal-viewid239.html>`. They consist in the heights reached in the two high tides as well as the heights reached in the two low tides. We have considered interval-valued tidal fluctuations determined by the minimum and the maximum heights that the tides reach in a day.

The purpose of the study is to determine if the variability of the tidal fluctuation in Gijón has changed in 2010 with respect to the one of the year 2009. We have information about the variability corresponding to the tides fluctuation in 2009 (given by interval data) taking into account the value

$\theta = 1/3$ (where $\theta = 1/3$ is the weight associated with the Lebesgue measure λ on $[0, 1]$ when the equivalence of Bertoluzza's metric and d_θ is considered, see [5, 24]). This choice of θ assigns less weight to the distances between the spreads (connected with the imprecision) than to the distances between the mids (connected with the location). The variability of the interval data provided in 2009 is $\sigma_0^2 = .0593$.

Note that in this case we are working with interval data and for this reason we are not going to use the old-fashioned real values' methodology but the intervals' methodology proposed in this work. Thus, if we consider the random interval $X \equiv$ *tidal fluctuation in Gijón in a day of 2010*, the idea is to solve the following test problem:

$$H_0 : \sigma_X^2 = .0593 \quad vs. \quad H_1 : \sigma_X^2 \neq .0593.$$

A simple random sample of size 50 has been drawn from X (since the bootstrap approach seems to have a good behaviour for a sample size equals to 50 in the two-sided test, as it is shown in Tables 3 and 4). These data are gathered in Table 5, and their sample variance is given by $\widehat{\sigma}_X^2 = .0663$. The asymptotic and bootstrap approaches proposed in this work have been applied (although the asymptotic approach is not really useful in this case due to the size of the sample considered), leading to p-values of $p = .4523$ in the asymptotic case and .4793 in the bootstrap one. Therefore, we can conclude that, at the most usually chosen significance levels, the variability of the tidal fluctuations in Gijón in 2010 cannot be considered to be different to the one in 2009, so there has not been big changes in the variability of the tides movement in 2010 with respect to the year 2009. As a consequence, we can consider that the tidal fluctuations in the Port of Gijón have more or less the same behaviour in the year 2010 than in the year 2009.

6 Conclusions

In this work, a methodology for analyzing the variability of the tidal fluctuations in a specific area has been proposed. This methodology is based on the use of intervals instead of real values for describing these fluctuations. In this context, some hypotheses tests concerning the variance of interval data have been presented which have been shown to be useful when moderate sample sizes (that is, sample sizes greater than or equal to 100) are involved.

It could be interesting to extend the proposed results to the case in which the variabilities in different areas are to be compared as well as some studies regarding the mean of the random intervals. In addition, a deeper sensitivity analysis can be carried out taking into account different choices for θ, as well as the distributions chosen to determine the intervals. Thus, the proposed methodology can be viewed as a starting point of a deeper analysis of the tidal fluctuations.

Table 5: Tidal fluctuations in Gijón in 2010

Day	Fluctuation	Day	Fluctuation	Day	Fluctuation	Day	Fluctuation
1	[.5,4.95]	14	[1, 4.41]	27	[.37, 5.03]	40	[.64, 4.87]
2	[1.62,3.86]	15	[1.86, 3.56]	28	[1.85, 3.56]	41	[.75, 4.74]
3	[1.15, 4.44]	16	[.94, 4.69]	29	[1.45, 4.08]	42	[1.14, 4.24]
4	[1.71, 3.58]	17	[1.79, 3.67]	30	[1.28, 4.09]	43	[1.24, 4.16]
5	[1.1, 4.31]	18	[1.21, 4.32]	31	[.39, 4.96]	44	[.63, 4.92]
6	[1.16, 4.4]	19	[1.46, 3.9]	32	[.85, 4.45]	45	[1.75, 3.72]
7	[.56, 4.75]	20	[1.46, 4.07]	33	[.97, 4.43]	46	[1.6, 3.91]
8	[1.6, 3.91]	21	[1.66, 3.71]	34	[1.38, 4.14]	47	[.73, 4.86]
9	[1.95, 3.3]	22	[1.15, 4.39]	35	[1.83, 3.46]	48	[.89, 4.41]
10	[1.32, 4.22]	23	[.96, 4.45]	36	[1.21, 4.04]	49	[1.74, 3.76]
11	[1.45, 3.82]	24	[1.33, 4.23]	37	[1.22, 4.34]	50	[1.88, 3.53]
12	[1.55, 3.67]	25	[1.83, 3.61]	38	[1.15, 4.11]		
13	[1.03, 4.36]	26	[1.13, 4.41]	39	[.78, 4.71]		

Acknowledgements The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grant MTM2009-09440-C02-02 and by the COST Action IC0702 (especially, through the two STSMs Ana Belén Ramos-Guajardo has been granted with). Their financial support is gratefully acknowledged.

References

1. Arstein Z, Vitale RA (1975) A strong law of large numbers for random compact sets. *Annals of Probability* 5:879–882
2. Aumann RJ (1965) Integrals of set-valued functions. *J. Math. Anal. Appl.* 12:1–12
3. Baker AC (1991) *Tidal power*. Peter Peregrinus Ltd., London
4. Beard AG, Mitchell NJ, Williams PJS, Kunitake M (1999) Non-linear interactions between tides and planetary waves resulting in periodic tidal variability. *Journal of Atmospheric and Solar-Terrestrial Physics* 61(5):363–376
5. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. *Mathware & Soft Computing* 2:71–84
6. Branzei R, Alparsan Gok SZ (2008) Bankuptcy problems with interval uncertainty. *Economics Bulletin* 3(56):1–10
7. Bryden IG, Couch SJ (2006) ME1-marine energy extraction: tidal resource analysis. *Renewable Energy* 31(2):133–139
8. Erskine AD (2005) The effect of tidal fluctuation on a coastal aquifer in the UK. *Ground Water* 29(4):556–562
9. Garrett C, Cummins P (2004) Generating Power from Tidal Currents. *Journal of Waterway, Port, Coastal, and Ocean Engineering* 130(3):114–118
10. Fréchet M (1948) Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* 10:215–310
11. Irpino A, Verde R (2008) Dynamic clustering of interval data using a Wasserstein-based distance. *Recognition Letters* 29(11):1648–1658

12. Gil MA, Lubiano MA, Montenegro M, López-García MT (2002) Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika* 56:97–111
13. Gil MA, Montenegro M, González-Rodríguez G, Colubi A, Casals MR (2006) Bootstrap approach to the multi-sample test of means with imprecise data. *Comput. Statist. Data Anal.* 51:148–162
14. González-Rodríguez G, Montenegro M, Colubi A, Gil MA (2006) Bootstrap techniques and fuzzy random variables: Synergy in hypothesis testing with fuzzy data. *Fuzzy Sets and Systems* 157:2608–2613
15. González-Rodríguez G, Blanco A, Corral N, Colubi A (2007) Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification* 1:67–81
16. Körner R (1997) *Linear models with random fuzzy variables*. PhD Thesis, Freiberg University of Mining and Technology
17. Kruse R, Meyer KD (1987) *Statistics with Vague Data*. D. Reidel Publishing Company, Dordrecht
18. Lubiano MA, Alonso C, Gil MA (1999) Statistical inferences on the S-mean squared dispersion of a fuzzy random variable. *Proc. Joint EUROFUSE-SIC99*, 532–537. Budapest
19. Montenegro M, Colubi A, Casals MR, Gil MA (2004) Asymptotic and Bootstrap techniques for testing the expected value of a fuzzy random variable. *Metrika* 59:31–49
20. Näther W (2000) On random fuzzy variables of second order and their application to linear statistical inference with fuzzy data. *Metrika* 51:201–221
21. Nio SD, Yang CS (1991) Sea-level fluctuations and the geometric variability of tide-dominated sandbodies. *Sedimentary Geology* 70(2–4):161–172, 179–193
22. Pipper CB, Ritz C (2006) Checking the grouped data version of the Cox Model for interval-grouped survival data. *Scandinavian Journal of Statistics* 34:405–418
23. Ramos-Guajardo AB, Colubi A, González-Rodríguez G, Gil MA (2010) One-sample tests for a generalized Fréchet variance of a fuzzy random variable. *Metrika* 71(2):185–202
24. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Information Sciences* 179:3964–3972

Comparing the Medians of a Random Interval Defined by Means of Two Different L^1 Metrics

Beatriz Sinova[1] and Stefan Van Aelst[2]

Abstract The standard central tendency measure for interval-valued data is the Aumann-type expected value, but as in real settings it is not always convenient because of the big influence that small changes in the data as well as the existence of great magnitude data have on its estimate. The aim of this paper is to explore other summary measures with a more robust behavior. The real-valued case has served as inspiration to define the median of a random interval. The definition of the median as a 'middle position' value is not possible here because of the lack of a universally accepted total order in the space of interval data, so the median is defined as the element which minimizes the mean distance, in terms of an L^1 metric (extension of the Euclidean distance in $\mathbb{R}$), to the values the random interval can take. The two metrics that we consider are the generalized Hausdorff metric (like the well-known Hausdorff metric, but including a positive parameter which determines the relative importance given to the difference in imprecision with respect to the difference in location) and the 1-norm metric introduced by Vitale. The aim of this paper is to compare these two approaches for the median of a random interval, both theoretically based on concepts commonly used in robustness and empirically by simulation.

1 Introduction and Motivation

Statistical data obtained from random experiments can be of a very different nature. Interval data frequently appear when intrinsically imprecise measurements (like fluctuations, ranges, censoring times, etc.) or values associated with some imprecise knowledge on numerical values (when dealing

[1] Universidad de Oviedo, Calvo Sotelo s/n, 33007 Oviedo, Spain, sinovabeatriz@uniovi.es
[2] Universiteit Gent, Krijgslaan 281, S9, B-9000 Gent, Belgium, Stefan.VanAelst@UGent.be

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 75–86.
springerlink.com

with grouped data for instance) are involved. Many examples can be found in real life, such as the intervals describing the age range covered by each class when individuals in surveys are split into age groups, the fluctuation of quotations on the stock exchange or the temperature range for the daily forecast in a certain location. Similarly, many interval data sets are obtained in research studies in different fields such as Medicine, Engineering, Empirical and Social Sciences in which the information about the range of values the variable takes along a period is even more relevant than the detailed records.

Random intervals are interval-valued random elements, that is, they formalize mathematically the random mechanism of producing interval data associated with a random experiment. To analyze this type of data, some central tendency measures based on the interval arithmetic (globally considering intervals as elements and not as sets of elements) have been proposed. The most often used measure is the Aumann-type expected value. It inherits very good probabilistic and statistical properties from the mean of a real-valued random variable, but that is also the reason why it can be highly influenced by the existence of great magnitude data or data changes.

In real settings, the solution is to consider more robust central tendency measures, like the median. Inspired by this, we define the median of a random interval. Taking into account that there is no universally accepted total order criterion in the space of non-empty compact intervals (so the median cannot be defined as a 'middle' position value), an L^1 metric, generalization of the Euclidean metric in $\mathbb{R}$, is required to define the median as the element of the space minimizing the mean distance to all the values the random interval can take. The first choice for the L^1 metric was the generalized Hausdorff metric (see Sinova *et al* [5]): a new distance based on the well-known Hausdorff metric expressed in terms of the mid/spr characterization of intervals (that is, their mid-point and their spread or radius). However, there are obstacles to generalize the median defined by means of the generalized Hausdorff metric to random fuzzy numbers due to the fact that, although the generalized mid and spread (see Trutschnig *et al.* [7]) characterize a fuzzy number, the sufficient and necessary conditions a function must fulfill to be a generalized mid or spread are not known yet and it is not possible to guarantee that the median defined in that way is indeed a fuzzy number. These difficulties prompted the use of another distance (suitable for the definition of the median of a random fuzzy number as shown in Sinova *et al.* [6]), based on the 1-norm, as introduced by Vitale [8], and which considers the characterization of an interval in terms of infima and suprema. Of course, a second definition of median of random intervals is obtained as a particular case of the median for random fuzzy numbers. The definition of both medians and their immediate properties are studied in Section 3, after recalling in Section 2 the notation and basic operations and concepts in the space of interval data. In Section 4, the two proposed definitions of median of a random interval are compared by means of the finite sample breakdown point and some simulation studies. Finally, Section 5 presents some conclusions and open problems.

2 The Space of Intervals $\mathcal{K}_c(\mathbb{R})$: Preliminaries

First of all, some notation is established, starting with $\mathcal{K}_c(\mathbb{R})$, the class of nonempty compact intervals. Each one of the intervals $K \in \mathcal{K}_c(\mathbb{R})$ can be characterized in terms of its infimum and supremum, $K = [\inf K, \sup K]$ or in terms of its mid-point and spread or radius, $K = [\operatorname{mid} K - \operatorname{spr} K, \operatorname{mid} K + \operatorname{spr} K]$, where

$$\operatorname{mid} K = \frac{\inf K + \sup K}{2}, \quad \operatorname{spr} K = \frac{\sup K - \inf K}{2}.$$

To analyze this kind of data, the two most relevant operations from a statistical point of view are the addition and the product by a scalar. In this paper, we use the usual interval arithmetic (the particular case of set arithmetic). That is:

- The *sum* of two nonempty compact intervals, $K, K' \in \mathcal{K}_c(\mathbb{R})$, is defined as the Minkowski sum of K and K', i.e., as the interval

$$K + K' = [\inf K + \inf K', \sup K + \sup K'] =$$

$$[(\operatorname{mid} K + \operatorname{mid} K') - (\operatorname{spr} K + \operatorname{spr} K'), (\operatorname{mid} K + \operatorname{mid} K') + (\operatorname{spr} K + \operatorname{spr} K')].$$

- The *product of an interval* $K \in \mathcal{K}_c(\mathbb{R})$ *by a scalar* $\gamma \in \mathbb{R}$ is defined as the element of $\mathcal{K}_c(\mathbb{R})$ such that

$$\gamma \cdot K = \begin{cases} [\gamma \cdot \inf K, \gamma \cdot \sup K] \text{ if } \gamma \geq 0 \\ [\gamma \cdot \sup K, \gamma \cdot \inf K] \text{ otherwise} \end{cases}$$

$$= [\gamma \cdot \operatorname{mid} K - |\gamma| \cdot \operatorname{spr} K, \gamma \cdot \operatorname{mid} K + |\gamma| \cdot \operatorname{spr} K].$$

A very important remark is that with these two operations the space is not linear, but only semilinear (with a conical structure) because of the lack of an opposite element for the Minkowski addition. Therefore, there is no generally applicable definition for the difference of intervals that preserves the connection with the sum in the real case. Hence, distances play a crucial role in statistical developments. Although L^2 metrics are very convenient in many statistical developments like least squares approaches, an L^1 distance is now needed in order to define the median. In this paper, the two following L^1 metrics will be used:

- The *generalized Hausdorff metric* (Sinova *et al.* [5]), which is partially inspired by the Hausdorff metric for intervals and the L^2 metrics in Trutschnig *et al.* [7]. It includes a positive parameter to weight the relative importance of the distance between the spreads relative to the distance between the mid-points (allocating the same weight to the deviation in location as to the deviation in imprecision is often viewed as a concern in the Hausdorff metric). Given two intervals $K, K' \in \mathcal{K}_c(\mathbb{R})$ and any

$\theta \in (0, \infty)$, the generalized Hausdorff metric between them is defined as:

$$d_{H,\theta}(K, K') = |\text{mid}\, K - \text{mid}\, K'| + \theta \cdot |\text{spr}\, K - \text{spr}\, K'|.$$

- The *1-norm metric*, introduced by Vitale [8]. Given any two intervals $K, K' \in \mathcal{K}_c(\mathbb{R})$, the 1-norm distance between them is:

$$\rho_1(K, K') = \frac{1}{2}|\inf K - \inf K'| + \frac{1}{2}|\sup K - \sup K'|.$$

  As mentioned before, this corresponds to the particular case (for intervals) of the metric used to define the median of random fuzzy numbers (Sinova *et al.* [5]).

A *random interval* is usually defined (following the random set-based approach to introduce this notion) as a Borel measurable mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R})$, starting from a probability space $(\Omega, \mathcal{A}, P)$, with respect to $\mathcal{A}$ and the Borel σ-field generated by the topology induced by the Hausdorff metric. The generalized Hausdorff metric and the 1-norm metric are topologically equivalent to each other and to the Hausdorff metric. Therefore, the definition of random interval can be rewritten in terms of either of these two metrics instead of the Hausdorff metric. This Borel measurability guarantees that concepts like the *distribution induced by a random interval* or the *stochastic independence of random intervals*, crucial for inferential developments, are well-defined by trivial induction. A random interval can also be defined in terms of real-valued random variables: X is a random interval if, and only if, both functions $\inf X : \Omega \to \mathcal{K}_c(\mathbb{R})$ and $\sup X : \Omega \to \mathcal{K}_c(\mathbb{R})$ (or equivalently, $\text{mid}\, X : \Omega \to \mathcal{K}_c(\mathbb{R})$ and $\text{spr}\, X : \Omega \to [0, \infty)$) are real-valued random variables.

The *Aumann expectation* is the standard central tendency measure for random intervals. This mean value is indeed the Fréchet expectation with respect to the d_θ metric, which corresponds to the Bertoluzza *et al.* [1] distance (see Gil *et al.* [3]) for the particular case of interval-valued data, and is defined as:

$$d_\theta(K, K') = \sqrt{(\text{mid}\, K - \text{mid}\, K')^2 + \theta \cdot (\text{spr}\, K - \text{spr}\, K')^2},$$

where $K, K' \in \mathcal{K}_c(\mathbb{R})$ and $\theta \in (0, \infty)$. This means that the Aumann expectation is the unique interval which minimizes, over $K \in \mathcal{K}_c(\mathbb{R})$, the expected squared distance $E[(d_\theta(X, K))^2]$. Furthermore, it can be expressed explicitly as the interval whose mid-point equals the expected value of $\text{mid}\, X$ and whose spread equals the expected value of $\text{spr}\, X$. The Aumann expectation inherits many very good probabilistic and statistical properties from the expectation of a real-valued random variable, like the linearity and invariance under linear transformations, and it also fulfills the Strong Law of Large Numbers for almost all the metrics we can consider. However, its high sensitivity to

data changes or extreme data makes this value not always convenient when summarizing the information given by interval-valued data sets.

3 The Median of a Random Interval Defined Through an L^1 Metric

The Aumann expectation of a random interval is not robust enough which is the motivation for extending the concept of median. Nevertheless, the non-existence of a universally accepted total order in the space $\mathcal{K}_c(\mathbb{R})$ does not allow us to define it as a 'middle position' value. In real settings another approach is to define the median as the value with the smallest mean Euclidean distance to the values of the real-valued random variable. Then, an L^1 metric between intervals which extends the Euclidean distance is required in order to define the median as the interval with the smallest mean distance to the values of the random interval. The two L^1 metrics between intervals introduced before satisfy this condition, so the definition of the median through both distances is now formalized.

Definition 1. The $d_{H,\theta}$-median (or medians) of a random interval $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ is (are) defined as the interval(s) $\mathrm{Me}[X] \in \mathcal{K}_c(\mathbb{R})$ such that:

$$E(d_{H,\theta}(X, \mathrm{Me}[X])) = \min_{K \in \mathcal{K}_c(\mathbb{R})} E(d_{H,\theta}(X, K)), \tag{1}$$

if these expected values exist.

A very practical result that guarantees the existence of the median and allows to compute it is the following. Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated random interval X, the minimization problem (1) has at least one solution, given by any nonempty compact interval such that:

$$\mathrm{mid}\ \mathrm{Me}[X] = \mathrm{Me}(\mathrm{mid}\, X), \quad \mathrm{spr}\ \mathrm{Me}[X] = \mathrm{Me}(\mathrm{spr}\, X).$$

It can immediately be noticed that the $d_{H,\theta}$-median is not unique if either $\mathrm{Me}(\mathrm{mid}\, X)$ or $\mathrm{Me}(\mathrm{spr}\, X)$ (which are medians of real-valued random variables) are not unique. It should be pointed out that the chosen solution does not depend on the value chosen for theta, although the mean error does.

Analogously, the median can be defined by means of the 1-norm metric:

Definition 2. The ρ_1-median (or medians) of a random interval $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ is (are) defined as the interval(s) $\mathrm{Med}[X] \in \mathcal{K}_c(\mathbb{R})$ such that:

$$E(\rho_1(X, \mathrm{Med}[X])) = \min_{K \in \mathcal{K}_c(\mathbb{R})} E(\rho_1(X, K)), \tag{2}$$

if these expected values exist.

In this situation, the practical choice (one of the solutions of minimization problem (2)) is the interval $\mathrm{Med}[X] \in \mathcal{K}_c(\mathbb{R})$ which satisfies:

$$\inf \mathrm{Med}[X] = \mathrm{Me}(\inf X), \quad \sup \mathrm{Med}[X] = \mathrm{Me}(\sup X).$$

If any of these two medians of real-valued random variables are not unique, the usual criterion of choosing the mid-point of the interval of possible medians is used to guarantee that $\mathrm{Med}(X)$ is nonempty.

Both medians preserve most of the elementary operational properties of the median in real settings. Namely,

Proposition 1. *Suppose that X is a random interval associated with a probability space. Then,*

- *if the distribution of X is degenerate at an interval value $K \in \mathcal{K}_c(\mathbb{R})$,*

$$\mathrm{Me}[X] = K,$$
$$\mathrm{Med}[X] = K.$$

- *for any $K \in \mathcal{K}_c(\mathbb{R})$ and $\gamma \in \mathbb{R}$,*

$$\mathrm{Me}[\gamma \cdot X + K] = \gamma \cdot \mathrm{Me}[X] + K,$$
$$\mathrm{Med}[\gamma \cdot X + K] = \gamma \cdot \mathrm{Med}[X] + K.$$

One remark about a distinctive feature in contrast to the real-valued case is that neither the $d_{H,\theta}$-median nor the ρ_1-median of a random interval is necessarily a value taken by the random interval as can be noticed from the following example: let X be a random interval taking the values $[0, 4]$, $[1, 3]$ and $[2, 5]$ with probability $\frac{1}{3}$. In this situation, the $d_{H,\theta}$-median is the interval $\mathrm{Me}[X] = [\mathrm{Me}(\mathrm{mid}\, X) - \mathrm{Me}(\mathrm{spr}\, X), \mathrm{Me}(\mathrm{mid}\, X) + \mathrm{Me}(\mathrm{spr}\, X)] = [2 - \frac{3}{2}, 2 + \frac{3}{2}] = [\frac{1}{2}, \frac{7}{2}]$ and the ρ_1-median is $\mathrm{Med}[X] = [\mathrm{Me}(\inf X), \mathrm{Me}(\sup X)] = [1, 4]$, neither of them being values the random interval takes.

As mentioned before, there is no universally accepted total order in the space $\mathcal{K}_c(\mathbb{R})$, so it is not possible to define the median as a 'middle position' value. However, both medians are a *measure of 'middle position'* with a certain partial ordering, when applicable. For the $d_{H,\theta}$-median, it can be proven that it is coherent with the Ishibuchi and Tanaka [4] partial ordering:

$$K \leq_{CW} K' \text{ if, and only if, } \mathrm{mid}\, K \leq \mathrm{mid}\, K' \text{ and } \mathrm{spr}\, K \geq \mathrm{spr}\, K'.$$

Hence, K' is considered to be *CW-larger* than K if, and only if, its location is greater and its imprecision is lower than for K:

Proposition 2. *For any sample of individuals $(\omega_1, \ldots, \omega_n)$ such that*

$$X(\omega_1) \leq_{CW} \ldots \leq_{CW} X(\omega_n)$$

we have that

- *if n is an odd number, then* $\mathrm{Me}[X] = X(\omega_{(n+1)/2})$,

- *if n is an even number, then* $\mathrm{Me}[X]$ = *any interval value 'between'* $X(\omega_{n/2})$ *and* $X(\omega_{(n/2)+1})$*, the 'between' being intended in the* $\leq_{CW}$ *sense, that is,* $\mathrm{mid}\, Me[X]$ *can be any value in* $\left[\mathrm{mid}\, X(\omega_{n/2}), \mathrm{mid}\, X(\omega_{(n/2)+1})\right]$*, whereas* $\mathrm{spr}\, Me[X]$ *can be any value in* $\left[\mathrm{spr}\, X(\omega_{(n/2)+1}), \mathrm{spr}\, (\omega_{n/2})\right]$.

On the other hand, the ρ_1-median is coherent with the well-known product order for the inf/sup vector, which is the partial ordering given by:

$$K \precsim K' \text{ if, and only if, } \inf K \leq \inf K' \text{ and } \sup K \geq \sup K'$$

or, equivalently, for all $\lambda \in [0,1]$ we have that $K^{[\lambda]} \leq K'^{[\lambda]}$, where $K^{[\lambda]} = \lambda \sup K + (1-\lambda) \inf K$.

Proposition 3. *For any sample of individuals* $(\omega_1, \ldots, \omega_n)$ *such that*

$$X(\omega_1) \precsim \ldots \precsim X(\omega_n)$$

we have that

- *if n is an odd number, then* $\mathrm{Med}[X] = X(\omega_{(n+1)/2})$,
- *if n is an even number, then* $\mathrm{Med}[X] = \dfrac{X(\omega_{n/2}) + X(\omega_{(n/2)+1})}{2}$.

Finally, the strong consistency of both the sample $d_{H,\theta}$-median and the sample ρ_1-median as estimators of the corresponding population quantities can be proven under very mild conditions as shown in the following results.

Proposition 4. *Suppose that X is a random interval associated with a probability space* $(\Omega, \mathcal{A}, P)$ *and* $\mathrm{Me}[X]$ *is unique. If* $\widehat{\mathrm{Me}[X]}_n$ *denotes the sample median associated with a simple random sample* $(X_1, \ldots, X_n)$ *from X, then*

$$\lim_{n\to\infty} d_{H,\theta}\left(\widehat{\mathrm{Me}[X]}_n, \mathrm{Me}[X]\right) = 0 \quad a.s.[P].$$

Proposition 5. *Suppose that X is a random interval associated with a probability space* $(\Omega, \mathcal{A}, P)$ *and* $\mathrm{Med}[X]$ *is unique without applying any convention. If* $\widehat{\mathrm{Med}[X]}_n$ *denotes the sample median associated with a simple random sample* $(X_1, \ldots, X_n)$ *from X, then*

$$\lim_{n\to\infty} \rho_1\left(\widehat{\mathrm{Med}[X]}_n, \mathrm{Med}[X]\right) = 0 \quad a.s.[P].$$

4 The Comparison between the $d_{H,\theta}$-median and the ρ_1-median of a Random Interval

The first result compares the $d_{H,\theta}$-median and the ρ_1-median by means of the computation of the finite sample breakdown point. Recall that the finite sample breakdown point is a measure of the robustness, since it gives the minimum proportion of sample data which should be arbitrarily increased or

decreased to make the estimate arbitrarily large or small. Following Donoho and Huber [2], the *finite sample breakdown point* (fsbp) of the sample $d_{H,\theta}$-median in a sample of size n from a random interval X is given by:

$$\text{fsbp}(\widehat{\text{Me}[X]}_n, x_n, d_{H,\theta})$$

$$= \frac{1}{n} \min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} d_{H,\theta}(\widehat{\text{Me}[X]}_n, \widehat{\text{Me}[Y_k]}_n) = \infty \right\},$$

where x_n denotes the considered sample of n data from the metric space $(\mathcal{K}_c(\mathbb{R}), d_{H,\theta})$ in which $\sup_{K,K' \in \mathcal{K}_c(\mathbb{R})} d_{H,\theta}(K, K') = \infty$ and $\text{Me}[Y_k]_n$ is the sample median of the sample $y_{n,k}$ obtained from the original sample x_n by perturbing at most k observations.

Analogously, the finite sample breakdown point of the sample ρ_1-median in a sample of size n from a random interval X is, with the same notation:

$$\text{fsbp}(\widehat{\text{Med}[X]}_n, x_n, \rho_1)$$

$$= \frac{1}{n} \min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} \rho_1(\widehat{\text{Med}[X]}_n, \widehat{\text{Med}[Y_k]}_n) = \infty \right\},$$

Then, it can be proven that

Proposition 6. *The finite sample breakdown point of both the sample $d_{H,\theta}$-median and the ρ_1-median from a random interval X, equal*

$$\text{fsbp}(\widehat{\text{Me}[X]}_n, x_n, d_{H,\theta}) = \text{fsbp}(\widehat{\text{Med}[X]}_n, x_n, \rho_1) = \frac{1}{n} \cdot \lfloor \frac{n+1}{2} \rfloor,$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

Proof. First note that the conditions $\sup_{K,K' \in \mathcal{K}_c(\mathbb{R})} d_{H,\theta}(K, K') = \infty$ and $\sup_{K,K' \in \mathcal{K}_c(\mathbb{R})} \rho_1(K, K') = \infty$ are fulfilled in the corresponding metric spaces because $d_{H,\theta}(\mathbf{1}_{[n-1,n+1]}, \mathbf{1}_{[-n-1,-n+1]}) = \rho_1(\mathbf{1}_{[n-1,n+1]}, \mathbf{1}_{[-n-1,-n+1]}) = 2n$. Since the fsbp for the sample median of a real-valued random variable equals $\lfloor \frac{n+1}{2} \rfloor$, we immediately have that:

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\text{mid}\,[X]}_n) - \text{Me}(\widehat{\text{mid}\,[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\text{spr}\,[X]}_n) - \text{Me}(\widehat{\text{spr}\,[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\inf[X]}_n) - \text{Me}(\widehat{\inf[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

$$\min \left\{ k \in \{1, \ldots, n\} : \sup_{y_{n,k}} |\text{Me}(\widehat{\sup[X]}_n) - \text{Me}(\widehat{\sup[Y_k]}_n)| = \infty \right\} = \lfloor \frac{n+1}{2} \rfloor$$

Therefore,

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} d_{H,\theta}(\widehat{\mathrm{Me}[X]}_n, \mathrm{Me}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n) \geq$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} |\mathrm{mid}\,(\widehat{\mathrm{Me}[X]}_n) - \mathrm{mid}\,(\mathrm{Me}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n)| =$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} |\mathrm{Me}(\widehat{\mathrm{mid}[\mathrm{X}]}_\mathrm{n}) - \mathrm{Me}(\widehat{\mathrm{mid}[\mathrm{Y}_{\lfloor\frac{n+1}{2}\rfloor}]}_\mathrm{n})| = \infty$$

and

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} \rho_1(\widehat{\mathrm{Med}[X]}_n, \mathrm{Med}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n) \geq$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} \frac{1}{2}|\inf(\widehat{\mathrm{Med}[X]}_n) - \inf(\mathrm{Med}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor}}]_n)|$$

$$= \frac{1}{2} \sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor}} |\mathrm{Me}(\widehat{\inf[X]}_n) - \mathrm{Me}(\widehat{\inf[Y_{\lfloor\frac{n+1}{2}\rfloor}]}_n)| = \infty$$

On the other hand,

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\mathrm{mid}\,[X]}_n) - \mathrm{Me}(\mathrm{mid}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_1 < \infty$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\mathrm{spr}\,[X]}_n) - \mathrm{Me}(\mathrm{spr}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_2 < \infty$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\inf[X]}_n) - \mathrm{Me}(\widehat{\inf[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_3 < \infty$$

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} |\mathrm{Me}(\widehat{\sup[X]}_n) - \mathrm{Me}(\widehat{\sup[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)| = M_4 < \infty$$

Consequently,

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} d_{H,\theta}(\widehat{\mathrm{Me}[X]}_n, \mathrm{Me}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor-1}}]_n)$$

$$= \sup_{Y_{\lfloor\frac{n+1}{2}\rfloor-1}} \Big[|\mathrm{Me}(\widehat{\mathrm{mid}[X]}_n) - \mathrm{Me}(\mathrm{mid}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)|$$

$$+ \theta \cdot |\mathrm{Me}(\widehat{\mathrm{spr}[X]}_n) - \mathrm{Me}(\mathrm{spr}\,\widehat{[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)|\Big] \leq M_1 + \theta \cdot M_2 < \infty$$

and

$$\sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} \rho_1(\widehat{\mathrm{Med}[X]}_n, \mathrm{Med}[\widehat{Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)$$

$$= \sup_{y_{n,\lfloor\frac{n+1}{2}\rfloor-1}} \left[\frac{1}{2} \cdot |\mathrm{Me}(\widehat{\inf[X]}_n) - \mathrm{Me}(\widehat{\inf[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)|\right.$$

$$+\frac{1}{2}\cdot\left|\mathrm{Me}(\widehat{\sup[X]}_n)-\mathrm{Me}(\widehat{\sup[Y_{\lfloor\frac{n+1}{2}\rfloor-1}]}_n)\right|\Big]\leq\frac{M_3+M_4}{2}<\infty$$

□

Furthermore, the fsbp of both medians can also be compared with the Aumann expectation:

Theorem 1. *The finite sample breakdown point of the sample Aumann expectation from a random interval* X, $\mathrm{fsbp}(\overline{X_n})$, *is lower than the ones for the sample* $d_{H,\theta}$*-median and the sample* ρ_1*-median for samples of size* $n>2$.

Proof. Following the same reasoning used in the previous proposition, it can be proven that

$$\mathrm{fsbp}(\overline{X_n},x_n,d_{H,\theta})=\mathrm{fsbp}(\overline{X_n},x_n,\rho_1)=\frac{1}{n},$$

so, consequently,

$$\mathrm{fsbp}(\widehat{\mathrm{Me}[X]}_n,x_n,d_{H,\theta})\geq\frac{n/2}{n}=\frac{1}{2}>\frac{1}{n}=\mathrm{fsbp}(\overline{X_n},x_n,d_{H,\theta})$$

$$\mathrm{fsbp}(\widehat{\mathrm{Med}[X]}_n,x_n,\rho_1)\geq\frac{n/2}{n}=\frac{1}{2}>\frac{1}{n}=\mathrm{fsbp}(\overline{X_n},x_n,\rho_1)$$

□

In order to corroborate these results, some empirical studies have been developed. A sample of $n=10000$ interval-valued data has been randomly generated from a random interval characterized by the distribution of two real-valued random variables, $\mathrm{mid}\,X$ and $\mathrm{spr}\,X$. Two cases have been considered: one in which the two random variables are independent (Case 1) and another one in which they are dependent (Case 2). In both situations, the sample has been split into two subsamples, one of size $n\cdot c_p$ associated with a contaminated distribution (hence c_p represents the proportion of contamination) and the other one, of size $n\cdot(1-c_p)$, without any perturbation. A second parameter, C_D, has also been included to measure the relative distance between the distribution of the contaminated and non contaminated subsamples. In detail, for different values of c_p and C_D the data for Case 1 are generated according to

- $\mathrm{mid}\,X\rightsquigarrow\mathcal{N}(0,1)$ and $\mathrm{spr}\,X\rightsquigarrow\chi^2_1$ for the non contaminated subsample,
- $\mathrm{mid}\,X\rightsquigarrow\mathcal{N}(0,3)+C_D$ and $\mathrm{spr}\,X\rightsquigarrow\chi^2_4+C_D$ for the contaminated subsample,

while for Case 2 we use

- $\mathrm{mid}\,X\rightsquigarrow\mathcal{N}(0,1)$ and $\mathrm{spr}\,X\rightsquigarrow\left(\frac{1}{(\mathrm{mid}\,X)^2+1}\right)^2+.1\cdot\chi^2_1$ for the non contaminated subsample,

- $\operatorname{mid} X \rightsquigarrow \mathcal{N}(0,3) + C_D$ and $\operatorname{spr} X \rightsquigarrow \left(\frac{1}{(\operatorname{mid} X)^2+1}\right)^2 + .1 \cdot \chi_1^2 + C_D$ for the contaminated subsample.

Both the population $d_{H,\theta}$-median and the population ρ_1-median are approximated by the Monte Carlo approach from this sample and the expected distance between the non contaminated distribution, X_{nc}, and the approximated medians, considering the $d_{H,\theta}$ and the ρ_1 distances, were computed.

c_p	c_D	Ratio_ρ	$\text{Ratio}_{\theta=1/3}$	$\text{Ratio}_{\theta=\sqrt{1/3}}$	$\text{Ratio}_{\theta=1}$	Ratio_ρ	$\text{Ratio}_{\theta=1/3}$	$\text{Ratio}_{\theta=\sqrt{1/3}}$	$\text{Ratio}_{\theta=1}$
.0	0	1.019406	1.010211	1.014805	1.020016	1.090363	1.071163	1.113693	1.173596
.0	1	1.019391	1.010212	1.014806	1.020017	1.090412	1.071170	1.113704	1.173612
.0	5	1.019393	1.010221	1.014805	1.020014	1.090448	1.071139	1.113654	1.173533
.0	10	1.019410	1.010209	1.014802	1.020012	1.090442	1.071171	1.113705	1.173613
.1	0	1.016934	1.008394	1.012000	1.015977	1.081155	1.066063	1.106141	1.163368
.1	1	1.017550	1.008663	1.012288	1.016226	1.072844	1.053439	1.085163	1.129607
.1	5	1.015010	1.007975	1.011077	1.014365	1.065343	1.046932	1.071485	1.102874
.1	10	1.011805	1.006462	1.008901	1.011478	1.046427	1.036585	1.054048	1.075179
.2	0	1.014011	1.006560	1.009286	1.012245	1.073723	1.061916	1.099925	1.154805
.2	1	1.014893	1.006741	1.009424	1.012272	1.056341	1.037449	1.059556	1.090469
.2	5	1.012616	1.006605	1.008862	1.011194	1.047532	1.028887	1.041835	1.057560
.2	10	1.009017	1.004951	1.006547	1.008209	1.029309	1.020252	1.028132	1.037146
.4	0	1.008012	1.003628	1.005115	1.006738	1.062304	1.055413	1.090023	1.140840
.4	1	1.006726	1.003075	1.004202	1.005429	1.024528	1.014247	1.022384	1.033863
.4	5	1.009291	1.007752	1.008795	1.009980	1.022742	1.014213	1.018332	1.022988
.4	10	1.006831	1.007307	1.008008	1.008899	1.012734	1.009485	1.011648	1.013964
.4	100	1.000904	1.000999	1.001095	1.001233	1.001385	1.001161	1.001371	1.001585

Table 1. Ratios of the mean distances of the mixed (partially contaminated and non-contaminated) sample $d_{H,\theta}$ and ρ_1-medians to the non-contaminated distribution of a random interval in Case 1 (left columns) and Case 2 (right columns)

In Table 1, the ratios $\text{Ratio}_\rho = E(\rho_1(X_{nc}, \text{Me}[X]))/E(\rho_1(X_{nc}, \text{Med}[X]))$ and $\text{Ratio}_\theta = E(d_{H,\theta}(X_{nc}, \text{Med}[X]))/E(d_{H,\theta}(X_{nc}, \text{Me}[X]))$ are shown. They show us how the mean distance increases (w.r.t. each metric) when the chosen median is not the one defined by means of the corresponding metric.

As Table 1 shows, the bigger the error proportion, the smaller the ratios. It can be also noticed that the smaller the θ, the smaller the corresponding ratio. As all the ratios are very close to 1, it can be concluded that both the $d_{H,\theta}$-median (with different choices for θ) and ρ_1-median have a quite similar behavior since there are no big differences between choosing one of the two measures in order to summarize the information given by the sample (independently from the distance used).

5 Concluding Remarks

In this study, two different definitions for the median of a random interval have been compared. Both definitions preserve important properties of the median in real settings and are coherent with the interpretation of the median as a 'middle position' value for certain partial orderings between intervals. By calculating the finite sample breakdown point and some simulation studies, the robustness of the two medians has been shown to be similar.

Future directions to be considered could be the extension of this comparison to the fuzzy-valued case and the definition of other central tendency measures. For instance, trimmed means or medians defined through depth functions could be adapted to this situation and compared with the current results.

Acknowledgements The research by Beatriz Sinova was partially supported by / benefited from the Spanish Ministry of Science and Innovation Grant MTM2009-09440-C02-01 and the COST Action IC0702. She has been also granted with the Ayuda del Programa de FPU AP2009-1197 from the Spanish Ministry of Education, an Ayuda de Investigación 2011 from the Fundación Banco Herrero and three Short Term Scientific Missions associated with the COST Action IC0702. The research by Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen). Their financial support is gratefully acknowledged.

References

1. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. *Mathware & Soft Comput.* 2:71–84
2. Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum K, Hodges Jr. JL (eds.) *A Festschrift for Erich L. Lehmann.* Wadsworth, Belmont
3. Gil MA, Lubiano MA, Montenegro M, López-García MT (2002) Least squares fitting of an affine function and strength of association for interval data. *Metrika* 56:97–111
4. Ishibuchi H, Tanaka H (1990) Multiobjective programming in optimization of the interval objective function. *Europ. J. Oper. Res.* 48:219–225
5. Sinova B, Casals MR, Colubi A, Gil MA (2010) The median of a random interval. In: Borgelt C, González-Rodríguez G, Trutschnig W, Lubiano MA, Gil MA, Grzegorzewski P, Hryniewicz O (eds.) *Combining Soft Computing and Statistical Methods in Data Analysis.* Springer, Heidelberg
6. Sinova B, Gil MA, Colubi A, Van Aelst S (2011) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets Syst.* in press (doi:10.1016/j.fss.2011.11.004)
7. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Inf. Sci.* 179:3964–3972
8. Vitale RA (1985) L_p metrics for compact, convex sets. *J. Approx. Theory* 45:280–287

Comparing the Representativeness of the 1-norm Median for Likert and Free-response Fuzzy Scales

Sara de la Rosa de Sáa[1] and Stefan Van Aelst[2]

Abstract Many questionnaires related to Social Sciences, Medical Diagnosis, Control Engineering, etc. are based on the well-known Likert scales. For its statistical data analysis each categorical response is usually encoded by an integer number. In this paper the convenience of allowing respondents to reply by using a free-response format based on the scale of fuzzy numbers is discussed by developing a comparative study through the mean 1-norm error on the representativeness of the corresponding median for the fuzzy and the integer-encoded Likert scales cases.

1 Introduction

Likert scales are widely employed in opinion/valuation/rating/... questionnaires which are usually associated with Social Sciences, Medical Diagnosis, Control Engineering, etc. They are often used for questionnaires with a prespecified response format, and the scale is easy to explain and to understand. Likert scale-based questionnaires involve several items on a topic and respondents should express their agreement/satisfaction/etc. by choosing one of k possible answers; k usually is in the range 3 to 10, although as argued by Lozano *et al.* [14] the most convenient choices from the psychometric and statistical viewpoint are in the range 4 to 7.

For the statistical analysis of the data, each of the responses is usually encoded by means of an integer number.

Several drawbacks of the statistical analysis of Likert responses have been pointed out in the literature. Among them we can outline the following:

[1] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, E-33007 Oviedo, Spain, sara16388@hotmail.com

[2] Department of Applied Mathematics and Computer Science, Ghent University, B-9000 Gent, Belgium, Stefan.VanAelst@UGent.be

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 87–98.
springerlink.com

- Likert scales discretize the variable associated with the response to a question into a small number of potential values;
- available techniques for the analysis of Likert scale data are rather limited and not very informative;
- integer encoding usually does not reflect the real differences between distinct 'values' (for instance, most of people consider the 'difference' between the responses encoded by 1 and 2 to be not the same as the difference between the responses encoded by 2 and 3).

To overcome these drawbacks some alternatives for the use of Likert scales have been suggested.

One of them, often referred to as the simple visual analogue scale, consists of choosing the single point within a concrete bounded interval/bar that 'best expresses' the response to the given question (cf. [1], [10], [15], [21]). This alternative involves a continuous scale that captures properly the diversity and relative variability in the response, and fits very well the statistical analysis. However, the choice of the point representing each response is not easy to make, and it does not seem realistic to demand such a level of accuracy in response to questions which are frequently intrinsically imprecise.

Another alternative is that of performing a fuzzification of the pre-specified Likert responses (cf. [9], [11], [12], [13], [20], [2, 3]). This alternative would involve a continuous scale as far as the support of the image is concerned, and a discrete one as far as the number of possible different values/responses is concerned. Although available data do not fit directly the best known statistical techniques, along the last decade a methodology for the statistical analysis of fuzzy data is being developed and many techniques have already been proposed and can be applied. Although the scale allows us to reflect the intrinsic imprecision of the Likert categories, the diversity and relative variability of the responses are not well-captured.

A third alternative is that of considering a tandem questionnaire format-scale integrating the positive features of the two first ones. In this respect, the scale of fuzzy numbers can be used in opinion, valuation, rating, etc. questionnaires to 'express' the responses, and these questionnaires could be designed with a free fuzzy response format. Advantages for this tandem are the following:

- the scale of fuzzy numbers is continuous as far as both the support of the image and the space of potential values are concerned;
- it is friendly-to-use and friendly-to-understand, and handling the scale does not require a very high expertise;
- it captures very well the diversity, relative variability and subjectivity in the response;
- it expresses very well the inherent imprecision of the categorical data;
- there are several well-developed techniques to analyze the responses from a statistical perspective.

Actually, the only critical point for this third alternative is shared with the two previous ones. This drawback concerns the practical conduction of these questionnaires in situations requiring a quick response (say on the street, at the door, by telephone, online, ...) without the required little time to explain respondents the 'fuzzy way' to reply.

This paper aims to discuss the interest of using the scale of fuzzy numbers-based questionnaires with a free response format versus the Likert scales questionnaires. In the paper this discussion is performed by developing a comparative study of the representativeness of the recently introduced 1-norm median (see Sinova *et al.* [17]) for the fuzzy data with the median of the integer-encoded Likert data.

Remark 1. It should be pointed out that the motivation for the study in this paper can be mostly found in the work by Hesketh and Hesketh along with collaborators (cf. [5, 6, 7, 8]). By combining the expertise in computing with that in psychometric evaluation, they have introduced the so-called *fuzzy rating scale* (and its computerized graphic version) as an extension of the semantic differential. This scale provides us with a common method of rating a variety of stimuli and analyzing/comparing the responses meaningfully across the stimuli. To ease the posterior analysis, Hesketh and Hesketh have asked the respondents to elicit their responses by using triangular fuzzy numbers for which the "$\vee$" pointer (upper vertex) means the preferred point and the left and right spreads indicate how far to the left or the right a particular rating can be possible/compatible (or, as Hesketh and Hesketh referred to, they determine the tolerable range of preferences). Fuzzy rating scales have been applied or considered to be potentially applicable to many areas such as job analysis, rating of selection interviews, performance ratings in customer services, and so on.

Hesketh and Hesketh outlined reasons for the interest of the statistical analysis of the collected responses. In fact, they have already performed some descriptive statistics with them, by analyzing separately the real-valued random variables associated with the three values characterizing triangular fuzzy numbers. Nowadays, the concept of random fuzzy set, and the developed statistical methodology would be certainly useful to draw much more conclusions from these responses and treat each datum as a whole. This methodology involves an added value: the possibility of drawing not only descriptive but also inferential conclusions. For this methodology responses don't need to be triangular, so that the preferred point in the fuzzy rating scale can be extended to be the class of values which are considered to be fully compatible with the respondent rating. The apparent computational complexity to develop statistics with the fuzzy responses can be substantially reduced by considering the R package SAFD (see [19] in this book and `http://cran.r-project.org/web/packages/SAFD/index.html`).

Remark 2. We would also like to remark that there is an essential difference between the approach for the fuzzy rating in this paper and the approaches

consisting of translating linguistic terms into fuzzy numbers. A noteworthy recent approach in the last respect has been developed by Bocklisch [3] and Bocklisch *et al.* [2]. This approach, motivated by the limitations of traditional statistics to deal with verbal response data, is based on a two-step procedure leading to an aggregate of the 'fuzzy translation' supplied by several participants in an empirical study. Unlike this approach, the one in this paper does not mean at all a translation converting the ordinal scale into a fuzzy numbered one, but it tries to take advantage of the richness of the fuzzy scale to avoid being constrained to a reduced number of different labels/values/ratings. In this way, each respondent draws (either by hand or computationally) his/her fuzzy numbered response specifically for each question posed and responses can be deeply refined (it is up to the respondent instead of up to a prefixed list).

2 Preliminaries

A *Likert scale-based questionnaire* corresponds to a survey in which for each question, the respondent is allowed to choose, among k answers, the one that best represents his/her opinion/valuation/rating/etc. Questionnaires based on Likert scales have (ordinal and frequently the same for each question) prespecified responses. Often, and especially for statistical purposes, responses are encoded by means of consecutive integer numbers (1-4, 1-5, 1-6, etc.). An example of a Likert scale-based question is given in Figure 1.

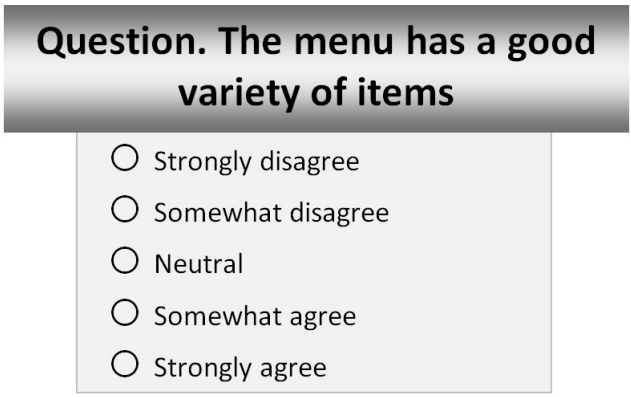


Fig. 1 Typical responses for a question concerning the variety of items in the menu of a restaurant

A (bounded) *fuzzy number* 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

for all $\alpha \in [0,1]$, the α-*level* set, $\widetilde{U}_\alpha = \{x \in \mathbb{R} \,|\, \widetilde{U}(x) \geq \alpha\}$ (for $\alpha > 0$) and $\widetilde{U}_0 = \mathrm{cl}\{x \in \mathbb{R} \,|\, \widetilde{U}(x) > 0\}$, is a nonempty compact interval. For each $x \in \mathbb{R}$, $\widetilde{U}(x)$ can be interpreted as the 'degree of compatibility' of x with $\widetilde{U}$. The space of bounded fuzzy numbers will be denoted by $\mathcal{F}_c^*(\mathbb{R})$.

In 'measuring' opinion, valuation, rating, etc., some strengths of the scale $\mathcal{F}_c^*(\mathbb{R})$ should be stressed, namely,

- it is much more expressive than an (ordinal) categorical scale and more realistic than its integer encoding;
- the transition from one value to another is gradual rather than abrupt;
- numerical/interval-valued data are special instances of fuzzy data;
- the usual arithmetic and distances between values in this scale pay attention either explicitly or implicitly to the 'location' and 'shape/imprecision' of values (which are crucial for their meaning and application).

Furthermore, an $\mathcal{F}_c^*(\mathbb{R})$-*based questionnaire* designed in accordance with a free response format allows us to capture appropriately the diversity, variability and subjectivity of the responses.

An example of an $\mathcal{F}_c^*(\mathbb{R})$-based response to the question in Figure 1 is displayed in Figure 2.

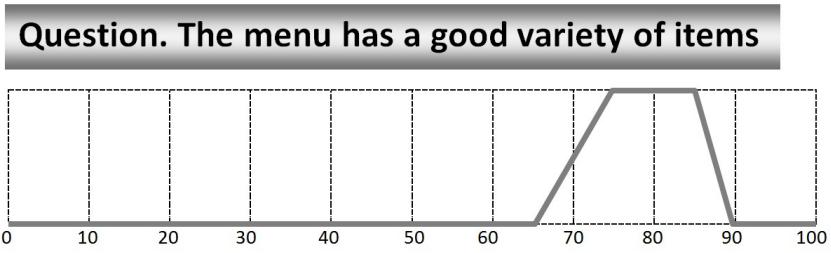


Fig. 2 Fuzzy (trapezoidal) response for a question concerning the variety of items in the menu of a restaurant

Concerning the statistical analysis of the collected responses, one can state that the analysis of Likert responses is usually performed by employing either techniques for categorical data or techniques for numerical data from variables with a small number of different integer values. In both approaches the number of applicable techniques is rather limited and relevant information associated with the diversity, relative variability and subjectivity, and even with the imprecise nature of the responses, is often lost or not enough exploited.

The analysis of fuzzy number valued responses is based on several key tools with specific features, such as the basic arithmetic, the measurement of the distance, and the probabilistic model for the random mechanism that generates the fuzzy responses along with the related summary measures.

In connection with the *arithmetic of fuzzy numbers*, the operations we should take into account are the sum and the product by scalars. These

operations are based on Zadeh's extension principle [22] which is equivalent to level-wise interval arithmetic. Thus, for $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c^*(\mathbb{R})$ and $\gamma \in \mathbb{R}$, $\widetilde{U} + \widetilde{V}$ is the fuzzy number such that for each $\alpha \in [0,1]$

$$(\widetilde{U} + \widetilde{V})_\alpha = \left[\inf \widetilde{U}_\alpha + \inf \widetilde{V}_\alpha, \sup \widetilde{U}_\alpha + \sup \widetilde{V}_\alpha\right]),$$

whereas $\gamma \cdot \widetilde{U}$ is the fuzzy number such that for each $\alpha \in [0,1]$

$$(\gamma \cdot \widetilde{U})_\alpha = \begin{cases} \left[\gamma \cdot \inf \widetilde{U}_\alpha, \gamma \cdot \sup \widetilde{U}_\alpha\right] & \text{if } \gamma \geq 0 \\ \left[\gamma \cdot \sup \widetilde{U}_\alpha, \gamma \cdot \inf \widetilde{U}_\alpha\right] & \text{if } \gamma < 0 \end{cases}$$

A first distinctive feature in contrast to the real-valued case lies in the fact that $(\mathcal{F}_c^*(\mathbb{R}), +, \cdot)$ does not have a linear structure (actually, it has a conical semilinear structure), since $\widetilde{U} + (-1) \cdot \widetilde{U} \neq \mathbf{1}_{\{0\}}$, the neutral element for the fuzzy sum. The semilinearity entails that one cannot establish a definition for the difference between fuzzy numbers that is well-defined and simultaneously preserves the properties of the difference between real numbers (the only way for these properties to be fulfilled is to consider the Hukuhara difference, which is not well-defined for most of the fuzzy numbers).

This inconvenience can often be overcome by using *metrics between fuzzy numbers* which are versatile and easy-to-use. Since the comparative analysis we present in this paper is based on the representativeness of the median, we consider the 1-norm metric by Diamond and Kloeden [4] which is an L^1 metric extending the Euclidean metric. For $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c^*(\mathbb{R})$ the 1-norm metric is defined as

$$\rho_1(\widetilde{U}, \widetilde{V}) = \frac{1}{2} \int_{(0,1]} \left(\left| \inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha \right| + \left| \sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha \right| \right) d\alpha.$$

The probabilistic model we consider to formalize the random mechanism generating the fuzzy numbered responses is that of *random fuzzy numbers* (RFNs, Puri and Ralescu [16], also called fuzzy random variables). Given a probability space $(\Omega, \mathcal{A}, P)$, a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R})$ is said to be a *random fuzzy number (RFN)* if for all $\alpha \in [0,1]$ the 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. Equivalently, an RFN can be formalized as a Borel-measurable mapping w.r.t. the Borel σ-field generated on $\mathcal{F}_c^*(\mathbb{R})$ by the topology associated with ρ_1. The Borel-measurability implies that one can properly refer to the distribution induced by an RFN, the statistical independence of RFNs, and so on. Hence we can refer adequately to a simple random sample from an RFN.

In Sinova *et al.* [17] the *1-norm median of an RFN* has been introduced as the fuzzy number $\widetilde{\mathrm{Me}}(\mathcal{X}) \in \mathcal{F}_c^*(\mathbb{R})$ such that

$$\left(\widetilde{\mathrm{Me}}(\mathcal{X})\right)_{\alpha} = \left[\mathrm{Me}\left(\inf \mathcal{X}_{\alpha}\right), \mathrm{Me}\left(\sup \mathcal{X}_{\alpha}\right)\right],$$

where in case $\mathrm{Me}\left(\inf \mathcal{X}_{\alpha}\right)$ or $\mathrm{Me}\left(\sup \mathcal{X}_{\alpha}\right)$ are nonunique we follow the most usual convention, i.e., we use the midpoint of the interval of medians. It can be proved that $\widetilde{\mathrm{Me}}$ satisfies many interesting properties such as

$$E\left(\rho_1\left(\mathcal{X}, \widetilde{\mathrm{Me}}(\mathcal{X})\right)\right) = \min_{\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R})} E\left(\rho_1\left(\mathcal{X}, \widetilde{U}\right)\right).$$

3 Empirical Comparative Study between Likert and Fuzzy Numbers Scales: The Mean (Absolute) Error Associated with the Median

To compare Likert and fuzzy scales from a statistical point of view, we examine the situation in which a person is simultaneously allowed to give a free response in the fuzzy scale, and to choose one of the possible responses in a Likert scale. Actually, we use a simulation design that mimics the human behavior. To generate the two types of responses, we first simulate fuzzy responses and then we 'Likertize' them by considering a plausible criterion.

Once data are simulated a crucial issue to think about is what we wish to compare. Many features can be chosen as suitable tools for statistical comparison. In this paper we consider the *representativeness of the median of the involved random element through the corresponding mean absolute error*.

The general simulation process is structured as follows: 1000 iterations of samples containing n trapezoidal fuzzy numbers are simulated ($n \in \{30, 50, 100, 300\}$); a trapezoidal fuzzy number $\widetilde{U}$ can be characterized as $\mathrm{Tra}(\inf \widetilde{U}_0, \inf \widetilde{U}_1, \sup \widetilde{U}_1, \sup \widetilde{U}_0)$. To generate each trapezoidal fuzzy response, we have followed the steps in Sinova *et al.* [18]. That is,

- A value $k \in \{4, 5, 6, 7\}$ is fixed.
- One value of the nonstandard (i.e., re-scaled and translated standard) beta distribution $(k-1) \cdot \beta(p, q) + 1$ is generated at random, with (p, q) varying to cover four different situations of distributions with values in $[1, k]$, namely, uniform, symmetrical weighting central values, symmetrical weighting extreme values, and an asymmetric one. The generated value is the mid-point of the 1-level, $\mathrm{mid}\, \widetilde{U}_1 = (\inf \widetilde{U}_1 + \sup \widetilde{U}_1)/2$.
- To avoid unusually 'wide' (and hence unrealistic) fuzzy responses, some constraints on the values for the deviations $\mathrm{mid}\, \widetilde{U}_1 - \inf \widetilde{U}_1 = \sup \widetilde{U}_1 - \mathrm{mid}\, \widetilde{U}_1$, $\inf \widetilde{U}_1 - \inf \widetilde{U}_0$ and $\sup \widetilde{U}_0 - \sup \widetilde{U}_1$ have been imposed, so that they have been generated from uniform distributions on the intervals $\left[0, \min\{(k-1)/10, \mathrm{mid}\, \widetilde{U}_1 - 1, k - \mathrm{mid}\, \widetilde{U}_1\}\right]$, $\left[0, \min\{k/5, \inf \widetilde{U}_1 - 1\}\right]$, and $\left[0, \min\{k/5, k - \sup \widetilde{U}_1\}\right]$, respectively; the trapezoidal fuzzy number is finally built from the generated mid-point and deviations.

Once the fuzzy responses are generated, they are 'Likertized'. It should be pointed out that the choice of the Likertization criterion seems not to be very relevant in this respect, although further research is needed to confirm this.

The chosen Likertization criterion is based on the metric ρ_1. If $\widetilde{U}$ is the generated fuzzy number, the criterion associates with it the integer number

$$\imath(\widetilde{U}) = \arg \min_{j\in\{1,\ldots,k\}} \rho_1(\widetilde{U}, \mathbf{1}_{\{j\}}).$$

As an illustration for this Likertization, we consider the fuzzy number in Figure 2 translated and re-scaled to interval [1,5] (i.e., we consider the trapezoidal fuzzy number $\widetilde{U} = \mathrm{Tra}(3.6, 4, 4.4, 6)$) and $k = 5$. We obtain that

$$\rho_1(\widetilde{U}, \mathbf{1}_{\{1\}}) = 3.15,\ \rho_1(\widetilde{U}, \mathbf{1}_{\{2\}}) = 2.15,\ \rho_1(\widetilde{U}, \mathbf{1}_{\{3\}}) = 1.15,$$

$$\rho_1(\widetilde{U}, \mathbf{1}_{\{4\}}) = 0.35,\ \rho_1(\widetilde{U}, \mathbf{1}_{\{5\}}) = 0.85,$$

whence $\imath(\widetilde{U}) = 4$.

The comparative analysis is based on examining the representativeness of the median in the encoded Likert and fuzzy scales. The mean absolute error (MAE) is considered to quantify this representativeness, where in the fuzzy case the metric ρ_1 is used to measure absolute errors.

The study has been performed for 4 different distributions for each of the 4 analyzed values of k. In each of these 4×4 cases 1000 samples of trapezoidal fuzzy numbers have been simulated for each considered sample of size n. For each sample we have computed:

- the FMAE(sample). If $\widetilde{x}_1, \ldots, \widetilde{x}_n$ are the values of $\mathcal{X}$ in the sample, then

$$\mathrm{FMAE(sample)} = \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\widetilde{x}_i, \widehat{\widetilde{\mathrm{Me}}}(\widetilde{x}_1, \ldots, \widetilde{x}_n) \right),$$

  where $\widehat{\widetilde{\mathrm{Me}}}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is the sample 1-norm median of $\widetilde{x}_1, \ldots, \widetilde{x}_n$ (or, equivalently, the 1-norm median of an RFN taking on values $\widetilde{x}_1, \ldots, \widetilde{x}_n$ with probabilities $1/n$),
- the LMAE(Lsample) (with Lsample = Likertized sample), that is,

$$\mathrm{LMAE(Lsample)} = \frac{1}{n} \sum_{i=1}^{n} \left(\imath(\widetilde{x}_i) - \widehat{\mathrm{Me}}(\imath(\widetilde{x}_1), \ldots, \imath(\widetilde{x}_n)) \right).$$

Moreover, along the 1000 samples we have computed the percentage of samples for which the FMAE(sample) was lower than the LMAE(Lsample).

Simulation results have been gathered in the following tables:

Table ($k = 4$). Simulations from $3 \cdot \beta(p, q) + 1$

(p, q)	n	% FMAE < LMAE
$(p, q) = (1, 1)$	30	70.7
	50	77.5
	100	90.3
	300	99.6
$(p, q) = (.75, .75)$	30	77.0
	50	86.0
	100	93.9
	300	99.9
$(p, q) = (1.1, 1.1)$	30	66.8
	50	77.9
	100	87.6
	300	98.7
$(p, q) = (6, 1)$	30	97.7
	50	99.1
	100	99.9
	300	100

Table ($k = 5$). Simulations from $4 \cdot \beta(p, q) + 1$

(p, q)	n	% FMAE < LMAE
$(p, q) = (1, 1)$	30	60.5
	50	56.8
	100	56.1
	300	50.0
$(p, q) = (.75, .75)$	30	73.0
	50	77.1
	100	80.5
	300	88.6
$(p, q) = (1.1, 1.1)$	30	56.5
	50	52.8
	100	47.6
	300	32.8
$(p, q) = (6, 1)$	30	98.6
	50	99.8
	100	100
	300	100

Table ($k = 6$). Simulations from $5 \cdot \beta(p, q) + 1$

(p, q)	n	% FMAE < LMAE
$(p, q) = (1, 1)$	30	60.5
	50	63.7
	100	73.6
	300	91.6
$(p, q) = (.75, .75)$	30	69.1
	50	76.1
	100	86.5
	300	96.4
$(p, q) = (1.1, 1.1)$	30	57.3
	50	59.8
	100	68.8
	300	87.6
$(p, q) = (6, 1)$	30	87.9
	50	94.6
	100	99.0
	300	100

Table ($k = 7$). Simulations from $6 \cdot \beta(p, q) + 1$

(p, q)	n	% FMAE < LMAE
$(p, q) = (1, 1)$	30	62.0
	50	61.4
	100	64.6
	300	64.5
$(p, q) = (.75, .75)$	30	70.7
	50	79.3
	100	83.9
	300	93.7
$(p, q) = (1.1, 1.1)$	30	58.1
	50	62.3
	100	58.9
	300	55.9
$(p, q) = (6, 1)$	30	71.3
	50	75.9
	100	83.3
	300	94.3

4 Concluding Remarks

The study in this paper is just an introductory one. The first empirical developments allow us to conclude that the use of the fuzzy versus Likert scales captures more diversity (this could be trivially and generally proved by using any diversity/entropy measure) and relative variability (this can be empirically proved by using some inequality indices). The simulations in the paper show that the median is in many, and often in most of the, situations more representative in fuzzy-free format than in the Likert scale questionnaire. This representativeness has been measured in terms of the error associated with 'estimating' each data by means of the median of all available ones. However, conclusions are not general enough in the last respect.

To enlarge and complement the study in the paper, immediate developments would be those formalizing by means of appropriate measures the fact that diversity and relative variability are much higher with fuzzy number free-format responses than with Likert ones. Also a deeper study should be performed for the representativeness of the Aumann-type mean value of an RFN, as well as to discuss some measures concerning the difference between the errors in both cases.

Furthermore, a sensitivity analysis should be carried out in connection with the choice of k, which is also a key term in psychometric studies on the use of Likert scales.

Acknowledgements The authors are grateful to the reviewers of the manuscript for their insightful comments and suggestions. This research has been partially supported by/benefited from the COST Action IC0702, the Spanish Ministry of Science and Innovation Grant MTM2009-09440-C02-01, a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen), and the Contract CP-PA-11-SMIRE from the Principality of Asturias-Universidad de Oviedo. Their financial support is gratefully acknowledged. De la Rosa de Sáa would like also thank the financial coverage of her Short Term Scientific Missions in Gent University through the COST Action.

References

1. Allen IE, Seaman CA (2007) Likert scales and Data Analyses. *Qual Prog* 40:64–65
2. Bocklisch FA, Bocklisch SF, Krems JF (2010) How to translate words into numbers? A fuzzy approach for the numerical translation of verbal probabilities. In: Hüllermeier E, Kruse R, Hoffmann F (eds) *Computational Intelligence for Knowledge-Based Systems Design*, LNAI 6178:614-623. Springer-Verlag, Heidelberg
3. Bocklisch FA (2011) The vagueness of verbal probability and frequency expressions. *Int J Adv Comp Sci* 1(2):52–57
4. Diamond P, Kloeden P (1999) Metric spaces of fuzzy sets. *Fuzzy Sets and Systems* 100:63–71
5. Hesketh B, Griffin B, Loh V (2011) A future-oriented retirement transition adjustment framework. *J Vocat Behav* 79:303–314

6. Hesketh B, Hesketh T, Hansen J-I, Goranson D (1995) Use of fuzzy variables in developing new scales from strong interest inventory. *J Couns Psych* 42:85–99
7. Hesketh T, Hesketh B (1994) Computerised fuzzy ratings: the concept of a fuzzy class. *Behav Res Meth, Inst & Comp* 26:272–277
8. Hesketh T, Pryor RGL, Hesketh B (1988) An application of a computerised fuzzy graphic rating scale to the psychological measurement of individual differences. *Int J Man Mach Stud* 29:21–35
9. Hu H-Y, Lee Y-C, Yen T-M (2010) Service quality gaps analysis based on Fuzzy linguistic SERVQUAL with a case study in hospital out-patient services. *The TQM J* 22:499–515
10. Kambaki-Vougioukli P, Vougiouklis, T (2008) Bar instead of scale. *Ratio Sociol* 3:49–56
11. Lalla M, Facchinetti G, Mastroleo G (2004) Ordinal scales and fuzzy set systems to measure agreement: an application to the evaluation of teaching activity. *Qual & Quant* 38:577–601
12. Lalla M, Facchinetti G, Mastroleo G (2008) Vagueness evaluation of the crisp output in a fuzzy inference system. *Fuzzy Sets and Systems* 159:3297–3312
13. Li CQ (2010) *A new Likert scale based on Fuzzy Set Theory.* PhD Thesis, Connecticut University, Connecticut, USA
14. Lozano LM, García-Cueto E, J. Muiz J (2008) Effect of the number of response categories on the reliability and validity of rating scales. *Methodology* 4:73-79
15. Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. *Adv in Health Sci Educ* 15:625–632
16. Puri ML, Ralescu DA (1986) Fuzzy random variables. *J Math Anal Appl* 114:409–422
17. Sinova B, Gil MA, Colubi A, Van Aelst S (2012) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets and Systems* In press:doi:10.1016/j.fss.2011.11.004
18. Sinova B, De la Rosa de Sáa S, Gil MA (2012) A generalized L^1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data, submitted
19. Trutschnig W, Lubiano MA, Lastra J (2012) SAFD — An R package for Statistical Analysis of Fuzzy Data. In: Borgelt C, Gil MA, Sousa JMC, Verleysen M (eds) *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, 107–118. Springer-Verlag, Heidelberg
20. Turksen IB, Willson IA (1994) A fuzzy set preference model for consumer choice. *Fuzzy Sets and Systems* 68:253–266
21. van Laerhoven H, van der Zaag-Loonen HJ, Derkx BHF (2004) A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta Pædiatr* 93:830–835
22. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Part 1. *Inform Sci* 8:199–249. Part 2. *Inform Sci* 8:301–353. Part 3. *Inform Sci* 9:43–80

Fuzzy Probability Distributions in Reliability Analysis, Fuzzy HPD-regions, and Fuzzy Predictive Distributions

Reinhard Viertl[1] and Shohreh Mirzaei Yeganeh[1]

Abstract In reliability analysis there are different kinds of uncertainty present: variability, imprecision of lifetimes, model uncertainty concerning probability distributions, and uncertainty of a-priori information in Bayesian analysis. For the description of imprecise lifetimes so-called fuzzy numbers are suitable. In order to model the uncertainty of a-priori information fuzzy probability distributions are the most up-to-date mathematical structure.

1 Introduction

The variability of lifetimes of similar units is usually described by random variables T and related probability distributions. More recently imprecise lifetime data, like lifetimes of trees in environmental statistics, recreation times after diseases, but also other data like amounts of toxic materials released to the environment, are modeled by so-called fuzzy numbers [1, 2]. These fuzzy numbers are special fuzzy subsets of the set of real numbers $\mathbb{R}$ whose membership functions $\xi(\cdot)$ sarisfy the following:

1. $\xi : \mathbb{R} \to [0; 1]$;
2. $\forall \delta \in (0; 1]$ the so-called δ-cut $C_\delta[\xi(\cdot)]$ defined by $C_\delta[\xi(\cdot)] := \{x \in \mathbb{R} : \xi(x) \geq \delta\} \neq \varnothing$, and all δ-cuts are finite unions of compact intervals;
3. $supp[\xi(\cdot)]$ is contained in a compact interval.

Functions $\xi(\cdot)$ fulfilling 1. - 3. are called *characterizing functions*. If all δ-cuts of a fuzzy number are compact intervals, then this fuzzy number is called *fuzzy interval*. The set of all fuzzy intervals is denoted by $\mathcal{F}_I(\mathbb{R})$.
A generalization of probability densities which is necessary in connection with

[1] Department of Statistics and Probability Theory, Vienna University of Technology, A-1040 Vienna, Austria, r.viertl@tuwien.ac.at · shohreh.mir@gmail.com

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 99–106.
springerlink.com

Bayesian inference for fuzzy data, are so-called *fuzzy probability densities* on measure spaces $(\mathcal{M}, \mathcal{A}, \mu)$.
A fuzzy probability density $f^*(\cdot)$ is a function $f^* : \mathcal{M} \to \mathcal{F}_I([0;\infty))$, i.e., a function whose values $f^*(x)$ are fuzzy intervals whose supports are subsets of the non-negative numbers $[0;\infty)$ for which all so-called *δ-level functions* $\underline{f}_\delta(.)$ and $\overline{f}_\delta(.)$ are integrable. These δ-level functions are defined by their values $\underline{f}_\delta(x)$ and $\overline{f}_\delta(x)$ by $C_\delta[f^*(x)] = [\underline{f}_\delta(x); \overline{f}_\delta(x)]$ for all $\delta \in (0;1]$ and all $x \in \mathcal{M}$. This means all integrals

$$\int_{\mathcal{M}} \underline{f}_\delta(x) d\mu(x) \quad \text{and} \quad \int_{\mathcal{M}} \overline{f}_\delta(x) d\mu(x)$$

exist and are finite. Based on fuzzy probability densities so-called *fuzzy probabilities of events* $A \in \mathcal{A}$ are determined as following.

The definition of fuzzy probabilities is based on a generating family of subsets of $\mathbb{R}$ to define a fuzzy interval via the so-called generation lemma for characterizing functions [3]. The generating intervals $[a_\delta; b_\delta]$ are defined using families $\mathcal{D}_\delta$ of classical probability densities $f(\cdot)$ on $(\mathcal{M}, \mathcal{A})$:

$$\mathcal{D}_\delta := \{f : \underline{f}_\delta(x) \le f(x) \le \overline{f}_\delta(x) \quad \forall x \in \mathcal{M}\}, \tag{1}$$

where a_δ and b_δ are defined by

$$a_\delta := \inf\{\int_A f(x) d\mu(x) : f \in \mathcal{D}_\delta\} \tag{2}$$

and

$$b_\delta := \sup\{\int_A f(x) d\mu(x) : f \in \mathcal{D}_\delta\} \qquad \forall \delta \in (0;1]. \tag{3}$$

The fuzzy probability $p^*(A)$ is the fuzzy interval whose characterizing function $\eta(\cdot)$ is given by

$$\eta(x) := \sup\{\delta \cdot \mathbf{1}_{[a_\delta; b_\delta]}(x) : \delta \in [0;1]\} \qquad \forall x \in \mathbb{R}, \tag{4}$$

where $\mathbf{1}_{[a_\delta;b_\delta]}(\cdot)$ is the indicator function of the interval $[a_\delta; b_\delta]$, and $[a_0; b_0] = \mathbb{R}$.

2 Fuzzy Lifetimes

In applied reliability analysis observed lifetimes as observations of time which is a continuous quantity are more or less fuzzy. Therefore a sample consists of n fuzzy numbers $t_1^*, \ldots, t_n^*$. The corresponding characterizing functions are denoted by $\xi_1(\cdot), \ldots, \xi_n(\cdot)$. Based on this kind of samples the reliability function $R(\cdot)$ can be estimated by a generalization of the *empirical reliability*

function (ERF) $\hat{R}_n(\cdot)$. For precise lifetimes $t_1, \ldots, t_n$ the ERF $\hat{R}_n(\cdot)$ is defined by its values

$$\hat{R}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(t;\infty)}(t_i) \qquad \forall t \geq 0. \tag{5}$$

In case of fuzzy lifetimes there are two possibilities to generalize $\hat{R}_n(\cdot)$. First a smoothed version of the ERF is obtained by

$$\hat{R}_n^{sm}(t) := \frac{1}{n} \sum_{i=1}^{n} \frac{\int_t^\infty \xi_i(x)dx}{\int_0^\infty \xi_i(x)dx} \qquad \forall t \geq 0. \tag{6}$$

The characterizing functions of four fuzzy lifetimes and the corresponding function $\hat{R}_n^{sm}(\cdot)$ are depicted in Figure 1. Alternatively a fuzzy valued generalization of the ERF, called *fuzzy empirical reliability function* (FERF), is obtained in the following way. Let the following functions $\hat{R}_{\delta,L}(\cdot)$ and $\hat{R}_{\delta,U}(\cdot)$ be defined for all $\delta \in (0;1]$ by

$$\left.\begin{aligned} \hat{R}_{\delta,U} &:= \tfrac{\#\{t_i^*:\ C_\delta(t_i^*) \cap (t;\infty) \neq \varnothing\}}{n} \\ \hat{R}_{\delta,L} &:= \tfrac{\#\{t_i^*:\ C_\delta(t_i^*) \subseteq (t;\infty)\}}{n} \end{aligned}\right\} \quad \forall t \geq 0. \tag{7}$$

From this definition the above functions are step functions fulfilling

$$\hat{R}_{\delta,L}(t) \leq \hat{R}_{\delta,U}(t) \qquad \forall t \geq 0. \tag{8}$$

Moreover $\hat{R}_{\delta,L}(0) = 1$ and $\hat{R}_{\delta,U}(0) = 1, \ \ \forall \delta \in (0;1]$ as well as

$$\lim_{t\to\infty} \hat{R}_{\delta,L}(t) = \lim_{t\to\infty} \hat{R}_{\delta,U}(t) = 0, \ \ \forall \delta \in (0;1]. \tag{9}$$

The FERF for the lifetime data in Figure 1 is depicted in Figure 2. For $\delta_1 < \delta_2$ the following holds true:

$$\hat{R}_{\delta_1,U}(t) \geq \hat{R}_{\delta_2,U}(t) \qquad \forall t \geq 0, \tag{10}$$

$$\hat{R}_{\delta_1,L}(t) \leq \hat{R}_{\delta_2,L}(t) \qquad \forall t \geq 0. \tag{11}$$

3 Bayesian Reliability Analysis

For parametric lifetime models $T \sim f(\cdot|\theta);\ \theta \in \Theta$ in Bayesian analysis also the parameter θ is described by a stochastic quantity $\tilde{\theta}$, whose probability distribution - before data are given - is called *a-priori distribution*. In case of continuous parameter space Θ the a-priori distribution is usually given by a probability density $\pi(\cdot)$ on Θ, called *a-priori density*. In case of precise data $t_1, \ldots, t_n$ the updating of the information concerning the distribution of the

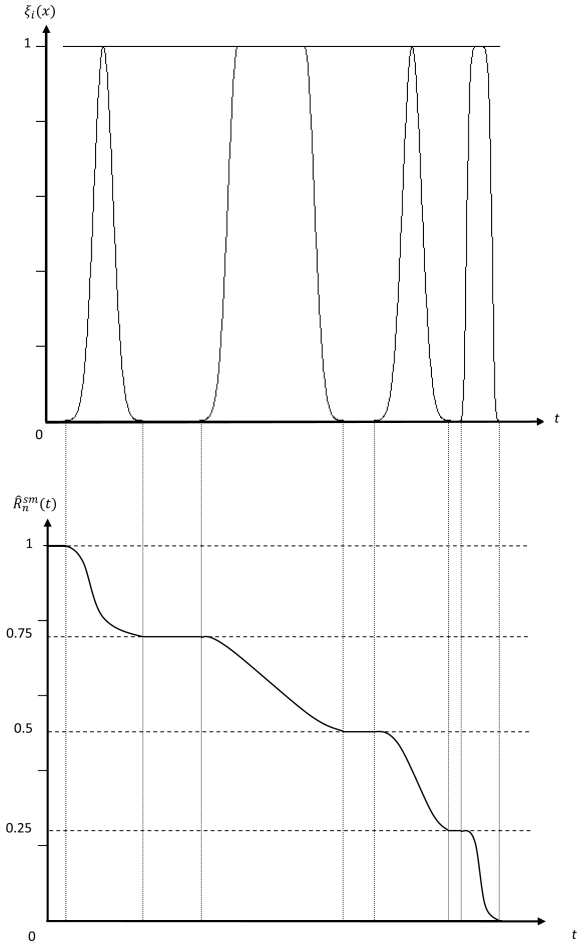


Fig. 1 Fuzzy lifetimes and smoothed empirical reliability function.

parameter is the so-called Bayes' theorem, i.e.,

$$\pi(\theta|t_1,\ldots,t_n) = \frac{\pi(\theta)\cdot l(\theta;t_1,\ldots,t_n)}{\int_\Theta \pi(\theta)\cdot l(\theta;t_1,\ldots,t_n)d\theta}, \qquad \forall\theta\in\Theta. \tag{12}$$

The conditional density $\pi(\cdot|t_1,\ldots,t_n)$ is called *a-posteriori density* of $\tilde{\theta}$. Based on the a-posteriori density *Bayesian confidence regions*, especially *HPD-regions*, as well as *predictive distributions* for lifetimes can be obtained. For fuzzy observed lifetimes $t_1^*,\ldots,t_n^*$ as described in Section 1, Bayes' theorem can be generalized in the following way.

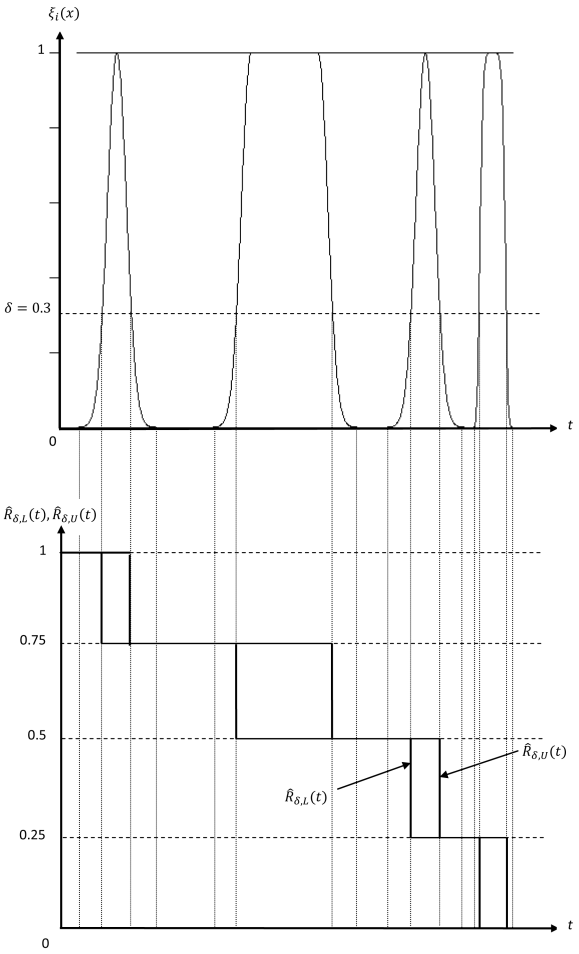


Fig. 2 Fuzzy sample and FERF for $\delta = 0.3$.

For continuous stochastic models $X \sim f(\cdot|\theta), \quad \theta \in \Theta$ with continuous parameter space Θ in general a-priori distributions as well as observations are fuzzy. Therefore it is necessary to generalize Bayes' theorem to this situation.

3.1 Likelihood Function for Fuzzy Data

In case of fuzzy data $t_1^*, \ldots, t_n^*$ the likelihood function $l(\theta; t_1, \ldots, t_n)$ has to be generalized to the situation of fuzzy numbers $t_1^*, \ldots, t_n^*$. The basis for that is the combined fuzzy sample element $\underline{t}^*$ [3]. Then the generalized likelihood

function $l^*(\theta; \underline{t}^*)$ is represented by its δ-level functions $\underline{l}_\delta(\cdot; \underline{t}^*)$ and $\overline{l}_\delta(\cdot; \underline{t}^*)$ for all $\delta \in (0; 1]$. For the δ-cuts of the fuzzy value $l^*(\theta; \underline{t}^*)$ we have

$$C_\delta(l^*(\theta; \underline{t}^*)) = [\underline{l}_\delta(\theta; \underline{t}^*), \overline{l}_\delta(\theta; \underline{t}^*)]. \tag{13}$$

Using this and the construction from [3] in order to keep the sequential property of the updating procedure in Bayes' theorem, the generalization of Bayes' theorem to the situation of fuzzy a-priori distribution and fuzzy data is possible.

Remark 1. The generalized likelihood function $l^*(\cdot; \underline{t}^*)$ is a fuzzy valued function, i.e., $l^* : \Theta \to \mathcal{F}_I([0, \infty))$.

3.2 Bayes' Theorem for Fuzzy A-priori Distribution and Fuzzy Data

Using the averaging procedure of δ-level curves of the a-priori density and combining it with the generalized likelihood function from Section 3.1, the generalization of Bayes' theorem is possible. The construction is based on δ-level functions.

Based on a fuzzy a-priori density $\pi^*(\cdot)$ on Θ with δ-level functions $\underline{\pi}_\delta(\cdot)$ and $\overline{\pi}_\delta(\cdot)$, and a fuzzy sample $t_1^*, \ldots, t_n^*$ with combined fuzzy sample $\underline{t}^*$ whose vector-characterizing function is $\zeta(\cdot, \ldots, \cdot)$, the characterizing function $\psi_{l^*(\theta; \underline{x}^*)}(\cdot)$ of $l^*(\theta; \underline{t}^*)$ is obtained by the extension principle, i.e.,

$$\psi_{l^*(\theta; \underline{x}^*)}(y) = \left\{ \begin{array}{ll} \sup\{\zeta(\underline{t}) : l(\theta; \underline{t}) = y\} & \text{if } \exists \underline{t} : l(\theta; \underline{t}) = y \\ 0 & \text{if } \nexists \underline{x} : l(\theta; \underline{t}) = y \end{array} \right\} \qquad \forall y \in \mathbb{R}. \tag{14}$$

The δ-level curves of the fuzzy a-posteriori density $\pi^*(\cdot | t_1^*, \ldots, t_n^*) = \pi^*(\cdot | \underline{t}^*)$ are defined in the following way:

$$\overline{\pi}_\delta(\theta | \underline{t}^*) := \frac{\overline{\pi}_\delta(\theta) \cdot \overline{l}_\delta(\theta; \underline{t}^*)}{\int_\Theta \frac{1}{2}[\underline{\pi}_\delta(\theta) \cdot \underline{l}_\delta(\theta; \underline{t}^*) + \overline{\pi}_\delta(\theta) \cdot \overline{l}_\delta(\theta; \underline{t}^*)] d\theta} \tag{15}$$

and

$$\underline{\pi}_\delta(\theta | \underline{t}^*) := \frac{\underline{\pi}_\delta(\theta) \cdot \underline{l}_\delta(\theta; \underline{t}^*)}{\int_\Theta \frac{1}{2}[\underline{\pi}_\delta(\theta) \cdot \underline{l}_\delta(\theta; \underline{t}^*) + \overline{\pi}_\delta(\theta) \cdot \overline{l}_\delta(\theta; \underline{t}^*)] d\theta} \tag{16}$$

for all $\delta \in (0; 1]$.

3.3 Generalized Fuzzy HPD-regions

From the a-posteriori density a generalization of confidence regions, especially *highest a-posteriori density regions* (HPD-regions) can be constructed.
Let $\pi^*(\cdot|t_1^*,\ldots,t_n^*)$ be the fuzzy a-posteriori density of $\tilde{\theta}$, and $\Theta \subseteq \mathbb{R}^k, \delta \in (0;1], \alpha \in (0;1), \alpha \ll 1$ and $1-\alpha$ the coverage probability. Moreover, defining $\mathcal{D}_\delta$ to be the set of classical probability densities g on Θ for which $\underline{\pi}_\delta(\theta) \leq g(\theta) \leq \overline{\pi}_\delta(\theta) \quad \forall \theta \in \Theta$, we define the generating system of subsets of Θ from which the generalized HPD-region, denoted as HPD*-region is obtained.
For $g \in \mathcal{D}_\delta$ let ${}^\delta HPD_{1-\alpha}(g)$ be the standard HPD-region for θ with coverage probability $1-\alpha$. Then the family of generating subsets of Θ, denoted by $(A_\delta; \delta \in (0;1])$, is defined by

$$A_\delta := \bigcup_{g \in \mathcal{D}_\delta} {}^\delta HPD_{1-\alpha}(g) \qquad \forall \delta \in (0;1]. \tag{17}$$

The membership function $\varphi(\cdot)$ of the HPD*-region is given by the so-called *construction lemma*, i.e.,

$$\varphi(\theta) := \sup\{\delta \cdot \mathbf{1}_{A_\delta}(\theta) : \delta \in [0;1]\} \qquad \forall \theta \in \Theta. \tag{18}$$

Remark 2. In case of classical a-posteriori density $\pi(\cdot|t_1,\ldots,t_n)$, the membership function $\varphi(\cdot)$ coincides with the indicator function $\mathbf{1}_{HPD_{1-\alpha}}(\cdot)$ of the classical HPD-region. This is seen by $\mathcal{D}_\delta = \{\pi(\cdot)\} \quad \forall \delta \in (0;1]$ and therefore ${}^\delta HPD_{1-\alpha} = HPD_{1-\alpha} \quad \forall \delta \in (0;1]$ which yields

$$\bigcup_{g \in \mathcal{D}_\delta} {}^\delta HPD_{1-\alpha}(g) = HPD_{1-\alpha} \qquad \forall \delta \in (0;1]. \tag{19}$$

Therefore $A_\delta = HPD_{1-\alpha} \quad \forall \delta \in (0;1]$, and $\varphi(\cdot) = \mathbf{1}_{HPD_{1-\alpha}}(\cdot)$.

3.4 Fuzzy Predictive Densities

Another application of fuzzy a-posteriori densities is the construction of generalized predictive densities $p(\cdot|t_1^*,\ldots,t_n^*)$ for lifetimes. In the classical case the predictive density is defined as the marginal density of the joint density of $(\tilde{\theta}, T)$, i.e.,

$$p(t|t_1,\ldots,t_n) = \int_\Theta f(t|\theta)\pi(\theta|t_1,\ldots,t_n)d\theta \qquad \forall t \geq 0. \tag{20}$$

In case of fuzzy a-posteriori densities $\pi^*(\cdot|t_1^*,\ldots,t_n^*)$ the above integral has to be generalized. This can be done in different ways [4]. The most suitable generalization seems to be the following: Again we look at $\mathcal{D}_\delta$ from above

and define for every $\delta \in (0;1]$ the closed interval $[a_\delta; b_\delta]$ by

$$b_\delta := \sup\{\int_\Theta f(x|\theta)g(\theta)d\theta : g \in \mathcal{D}_\delta\} \tag{21}$$

$$a_\delta := \inf\{\int_\Theta f(x|\theta)g(\theta)d\theta : g \in \mathcal{D}_\delta\}. \tag{22}$$

The characterizing function $\psi_t(\cdot)$ of the value $p^*(t|t_1^*, \ldots, t_n^*) \quad \forall t \geq 0$ of the generalized fuzzy predictive density $p^*(\cdot|t_1^*, \ldots, t_n^*)$ is defined by the construction lemma:

$$\psi_t(y) := \sup\{\delta \cdot \mathbf{1}_{[a_\delta;b_\delta]}(y) : \delta \in [0;1]\} \qquad \forall y \in \mathbb{R} \tag{23}$$

Remark 3. For precise a-posteriori density the result coincides with the result from standard Bayesian inference.

4 Conclusion

Fuzzy observed lifetimes make it necessary to generalize the methods of reliability analysis to this kind of data. In Bayesian reliability analysis, the corresponding a-posteriori distributions become fuzzy. Therefore the consideration of fuzzy probability distributions is necessary. Based on that a generalization of the concept of highest a-posteriori density regions for parameters is given in this paper as well as a generalization of predictive densities for lifetimes based on fuzzy data.

References

1. Möller B and Beer M (2004) *Fuzzy Randomness — Uncertainty in Civil Engineering and Computational Mechanics.* Springer, Berlin
2. Viertl R (2009) On reliability estimation based on fuzzy lifetime data. *Journal of Statistical Planning and Inference* 139:1750–1755
3. Viertl R (2011) *Statistical Methods for Fuzzy Data.* Wiley, Chichester
4. Viertl R (2011) On predictive densities in fuzzy Bayesian inference. In: Faber M et al. (eds.): *Applications of Statistics and Probability in Civil Engineering.* Taylor & Francis, London

SAFD — An R Package for Statistical Analysis of Fuzzy Data

Wolfgang Trutschnig[1], María Asunción Lubiano[2], and Julia Lastra[1]

Abstract The R package SAFD (Statistical Analysis of Fuzzy Data) provides basic tools for elementary statistics with one dimensional Fuzzy Data in the form of polygonal fuzzy numbers. In particular, the package contains functions for the standard operations on the class of fuzzy numbers (sum, scalar product, mean, Hukuhara difference, quantiles) as well as for calculating (Bertoluzza) distance, sample variance, sample covariance, sample correlation, and the Dempster-Shafer (levelwise) histogram. Moreover SAFD facilitates functions for the simulation of fuzzy random variables, for bootstrap tests for the equality of means as well as a function for linear regression given trapezoidal fuzzy data. The aim of this paper is to explain the functionality of the package and to illustrate its usage by various examples.

1 Introduction

During the last decades the concept of fuzzy sets going back to L. Zadeh (see [13]) has become more and more popular, particularly in order to model imprecision that typically arises in the context of collecting or processing different kinds of realistic data. Firstly, as a matter of fact, all measurements of continuous physical quantities are imprecise - the imprecision may, for instance, be caused by the fact that the measurement device rounds to a certain number of digits or may be due to physical conditions (Heisenberg uncertainty in the microscopic level, etc.). And secondly, humans typically quantify and classify using linguistic labels, which only in very rare cases can properly be modelled by exact (crisp) real values. It is therefore not sur-

[1] European Centre for Soft Computing, Edificio de Investigación, Calle Gonzalo Gutiérrez Quirós s/n, 33600 Mieres (Asturias), Spain, wolfgang.trutschnig@softcomputing.es
[2] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, 33007 Oviedo (Asturias), Spain, lubiano@uniovi.es

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 107–118.
springerlink.com

prising that fuzzy set theory has been applied in data analysis problems in various areas like forestry, structural analysis, hydrology and economics (see [2, 3, 5, 7]). Combining probabilistic uncertainty with (one-dimensional) imprecision naturally leads to the concept of so-called fuzzy random variables, which are random elements X on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ with values in the family $\mathcal{F}_c(\mathbb{R})$ of all fuzzy numbers (see [8]). Even from the purely descriptive point of view one needs some software that allows to analyze and plot samples of fuzzy numbers (from whatever source they may come). From the inferential point of view this software certainly should also be able to generate fuzzy samples, i.e. to simulate fuzzy random variables.
Since it seemed most natural to write this software not as a stand-alone product but as add-on to an already existing, world wide used, state-of-the art and sufficiently broad software environment we have decided to write an R package entitled SAFD (Statistical Analysis of Fuzzy Data). R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS and is available under http://www.r-project.org/. Currently, SAFD is one of more than 3.000 packages on CRAN, see http://cran.r-project.org/web/packages/.
In its current version 0.3 the package contains functions that allow to execute standard operations on the class $\mathcal{F}_c(\mathbb{R})$ (sum, scalar product, mean, Hukuhara difference, quantiles) as well as functions calculating (Bertoluzza) distance, sample variance, sample covariance, sample correlation, and the Dempster-Shafer (levelwise) histogram. Moreover SAFD facilitates functions to simulate fuzzy random variables, bootstrap tests for the equality of means as well as a function for linear regression given trapezoidal fuzzy data.
The aim of this contribution is to explain the functionality of the package and to illustrate its usage by various examples - hence the rest of the paper is organized as follows: Firstly some preliminaries and notations concerning fuzzy numbers and fuzzy random variables are gathered in Section 2. Section 3 explains the most important functions with the help of various examples. Finally, Section 4 presents some possible future extensions for SAFD.

Remark: In its current version 0.3 SAFD is still a quite small package that should be extended in the near future. Any recommendations, comments and complaints are welcome.

2 Notation and Preliminaries

Throughout the whole paper $\mathcal{K}_c(\mathbb{R})$ denotes the family of all non-empty compact intervals. The family of all *fuzzy numbers* $\mathcal{F}_c(\mathbb{R})$ considered throughout this paper is defined by

$$\mathcal{F}_c(\mathbb{R}) = \left\{ A : \mathbb{R} \to [0,1] \;\middle|\; A_\alpha \in \mathcal{K}_c(\mathbb{R}) \;\text{ for every } \alpha \in [0,1] \right\}, \qquad (1)$$

whereby the α-cut A_α is defined by

$$A_\alpha := [\underline{a}_\alpha, \overline{a}_\alpha] := \{x \in \mathbb{R} \,|\, A(x) \geq \alpha\}$$

for every $\alpha \in (0,1]$ and the 0-cut $A_0 := [\underline{a}_0, \overline{a}_0]$, called the support of A, is defined as the topological closure of $\bigcup_{\alpha>0} A_\alpha$. For every pair $A, B \in \mathcal{F}_c(\mathbb{R})$ and $b \in \mathbb{R}$ the Minkowski sum $S = A \oplus B \in \mathcal{F}_c(\mathbb{R})$ of A and B and the scalar product $P = b \odot A \in \mathcal{F}_c(\mathbb{R})$ are levelwise defined by

$$[\underline{s}_\alpha, \overline{s}_\alpha] = [\underline{a}_\alpha + \underline{b}_\alpha, \overline{a}_\alpha + \overline{b}_\alpha] \quad \forall \alpha \in [0,1]$$

and

$$[\underline{p}_\alpha, \overline{p}_\alpha] = [b\,\underline{a}_\alpha, b\,\overline{a}_\alpha] \quad \forall \alpha \in [0,1]$$

if $b \geq 0$ and

$$[\underline{p}_\alpha, \overline{p}_\alpha] = [b\,\overline{a}_\alpha, b\,\underline{a}_\alpha] \quad \forall \alpha \in [0,1]$$

if $b \leq 0$. If, for given $A, B \in \mathcal{F}_c(\mathbb{R})$ there is a fuzzy number $D \in \mathcal{F}_c(\mathbb{R})$ such that $A = B \oplus D$ then D will be called *Hukuhara* difference of A and B and denoted by $A \ominus_H B$. Note that, in case it exists, the Hukuhara difference is the inverse operation to $\oplus$.

In order to measure distances between elements in $\mathcal{F}_c(\mathbb{R})$ we will consider the simplest form of the so-called *Bertoluzza metric* d_θ with $\theta > 0$ (see [1] and [11]), which is defined by

$$d_\theta^2(A,B) := \int_{[0,1]} (mid(A_\alpha) - mid(B_\alpha))^2 \, d\alpha + \theta \int_{[0,1]} (spr(A_\alpha) - spr(B_\alpha))^2 \, d\alpha, \tag{2}$$

whereby $mid(A_\alpha) := \frac{1}{2}(\underline{a}_\alpha + \overline{a}_\alpha)$ denotes the midpoint of A_α and $spr(a_\alpha) := \frac{1}{2}(\overline{a}_\alpha - \underline{a}_\alpha)$ the spread (or radius) of A_α for every $\alpha \in [0,1]$.

Throughout this paper a *fuzzy random variable* is a Borel measurable mapping X from a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ into the metric space $(\mathcal{F}_c(\mathbb{R}), d_\theta)$. Note that Borel measurability does not depend on the concrete choice of θ in the definition of d_θ as long as $\theta > 0$ since all metrics d_θ induce the same topology on $\mathcal{F}_c(\mathbb{R})$.
Given a sample of a fuzzy random variable X, $\mathfrak{X} = (X_1, \cdots, X_n)$, the sample mean $\overline{X}_n$ and sample variance $\sigma^2_{\theta,n}$ are defined by

$$\overline{X}_n := \frac{1}{n}(X_1 \oplus X_2 \oplus \cdots \oplus X_n) \tag{3}$$

and

$$\sigma^2_{\theta,n} := \frac{1}{n} \sum_{i=1}^{n} d_\theta^2(X_i, \overline{X}_n). \tag{4}$$

Furthermore, if $\mathfrak{X} = (X_1, \cdots, X_n)$ and $\mathfrak{Y} = (Y_1, \cdots, Y_n)$ are samples of the fuzzy random variables X and Y, their (Bertoluzza) covariance (see [3]) is defined by

$$cov_\theta(\mathfrak{X}, \mathfrak{Y}) := cov_{mid}(\mathfrak{X}, \mathfrak{Y}) + \theta\, cov_{spr}(\mathfrak{X}, \mathfrak{Y}),$$

whereby $cov_{mid}(\mathfrak{X}, \mathfrak{Y})$ is the (integral) mean sample covariance of the corresponding mids and $cov_{spr}(\mathfrak{X}, \mathfrak{Y})$ the (integral) mean sample covariance of the corresponding spreads, i.e.

$$\begin{aligned} cov_{mid}(\mathfrak{X}, \mathfrak{Y}) &= \int_{[0,1]} \frac{1}{n} \sum_{i=1}^{n} mid((X_i)_\alpha)\, mid((Y_i)_\alpha) d\alpha \\ &\quad - \int_{[0,1]} mid((\overline{X}_n)_\alpha)\, mid((\overline{Y}_n)_\alpha) d\alpha \\ cov_{spr}(\mathfrak{X}, \mathfrak{Y}) &= \int_{[0,1]} \frac{1}{n} \sum_{i=1}^{n} spr((X_i)_\alpha)\, spr((Y_i)_\alpha) d\alpha \\ &\quad - \int_{[0,1]} spr((\overline{X}_n)_\alpha)\, spr((\overline{Y}_n)_\alpha) d\alpha. \end{aligned}$$

For every sample $\mathfrak{X} = (X_1, X_2 \ldots, X_n)$ of fuzzy numbers we will consider levelwise quantiles, so, for instance, the median $med(\mathfrak{X})$ is defined as the unique element of $\mathcal{F}_c(\mathbb{R})$ such that for all $\alpha \in [0,1]$ we have

$$med(\mathfrak{X})_\alpha = \left[\overline{F}^{-1}_{n,\alpha}(0.5), \underline{F}^{-1}_{n,\alpha}(0.5)\right], \tag{5}$$

whereby the functions $\overline{F}^{-1}_{n,\alpha}$ and $\underline{F}^{-1}_{n,\alpha}$ denote the (pseudo-) inverse of the empirical distribution function of the sample $(\underline{x_1}_\alpha, \underline{x_2}_\alpha, \cdots \underline{x_n}_\alpha)$ and the sample $(\overline{x_1}_\alpha, \overline{x_2}_\alpha, \cdots \overline{x_n}_\alpha)$ respectively (also see [9] for a slightly different definition).

Suppose that $I \in \mathcal{K}_c(\mathbb{R})$ and that $A_1, A_2 \ldots, A_n \in \mathcal{F}_c(\mathbb{R})$, then the fuzzy relative frequency of I is defined as the unique fuzzy number $H_n(I) \in \mathcal{F}_c(\mathbb{R})$ such that all but at most countably many α-cuts $(H_n(I))_\alpha := [\underline{h}_{n,\alpha}(I), \overline{h}_{n,\alpha}(I)]$ fulfil

$$\begin{aligned} \underline{h}_{n,\alpha}(I) &= \frac{1}{n} \#\Big\{ i \in \{1, \ldots, n\} : (A_i)_\alpha \subseteq I \Big\} \\ \overline{h}_{n,\alpha}(I) &= \frac{1}{n} \#\Big\{ i \in \{1, \ldots, n\} : (A_i)_\alpha \cap I \neq \emptyset \Big\}. \end{aligned} \tag{6}$$

In other words, all but at most countably many α-cuts of $H_n(I)$ coincide with Dempster's interval-valued frequency of the corresponding α-cuts of the sample. For a more detailed and more general description of fuzzy frequencies please see [10] and [12].

3 Functionality of the SAFD Package

The SAFD package works with *polygonal fuzzy numbers* in the form of *dataframes* having two columns, `x` and `alpha`, and arbitrary many α-levels. The following is an example with only three equidistant levels:

```
> A
        x alpha
1 -2.2000   0.0
2 -1.2000   0.5
3 -0.2000   1.0
4  0.2000   1.0
5  1.1375   0.5
6  3.2000   0.0
```

Thereby the x-values of the dataframe have to be increasing and the alpha-values have to increase from 0 to 1 and then decrease from 1 to 0 in the same manner (non necessarily equidistant). The package contains two internal functions called `checking` and `checking2` to check if the data is in the correct format. Furthermore it contains a function called `translator` to convert input data fulfilling the above conditions into a dataframe in the correct format and with a chosen number `nl` of equidistant alpha levels, see Figure 1.

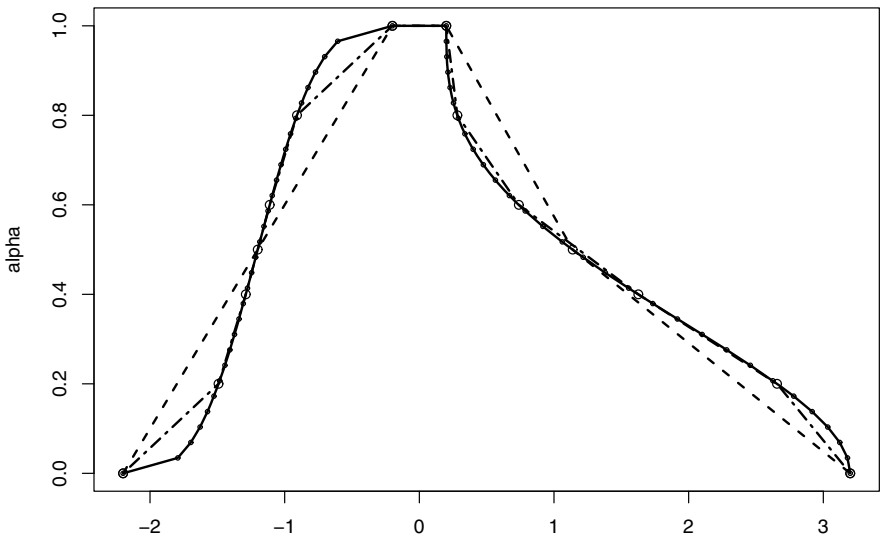


Fig. 1 Illustration of the functionality of the `translator` function via the following commands (the XX data set is included in the package):
`data(XX); E3<-translator(XX[[3]],3); E6<-translator(XX[[3]],6)`

Samples of polygonal fuzzy numbers are handled as *lists of dataframes* in the above format, i.e. XX[[3]] is the third fuzzy number of the sample XX.
In the following we start with some very simple examples illustrating how to use the most basic functions contained in the package. These basic functions are also used by the other functions in the package like `Bvar`, `btest*.mean` or `lrmodel`. Further examples for these functions are given in the html help files included in SAFD. Afterwards we take a look to some more interesting examples involving the generation of samples and the calculating and plotting of frequencies/histograms for the generated samples in Section 3.2.

3.1 Basic Functions

Given polygonal fuzzy numbers $X_1, \cdots, X_n$ contained in a list `AA` their *Minkowski mean* $\overline{X}_n$ can be calculated as illustrated in the following example, which also shows how to define polygonal fuzzy numbers by hand. Additionally we can easily calculate the Bertoluzza distance $d_{1/3}(X, Y)$ for X, Y from Example 1[1]. The results produced by Example 1 are depicted in Figure 2.

Example 1:

```
X<-translator(XX[[3]],10)
E<-data.frame(cbind(x=c(2,3,3.5,4),alpha=c(0,1,1,0)))
Y<-translator(E,10)
AA<-list(X,Y)
M<-Mmean(AA, pic=1)
bertoluzza(X,Y,theta=1/3,pic=1)
```

For examples concerning the functions `Msum` and `sc_mult` (which are also very basic functions) we refer to the html help of the SAFD package and only remark here that most functions in the package operate on families of polygonal fuzzy numbers having *identical* α-levels in order to allow the operations to be executed as quickly as possible[2]. Using the `translator` function this is, however, no real restriction since all elements of interest can easily be transformed in the correct form.

3.2 Sample Generation, Frequency and Histogram

The SAFD package contains a function called `generator` that allows to simulating fuzzy random variables. This function is an implementation of the

[1] 1/3 is the default value for θ for all functions involving the Bertoluzza distance

[2] identical α-levels allow, for instance, to calculate the sum of polygonal fuzzy numbers via elementary vector arithmetic

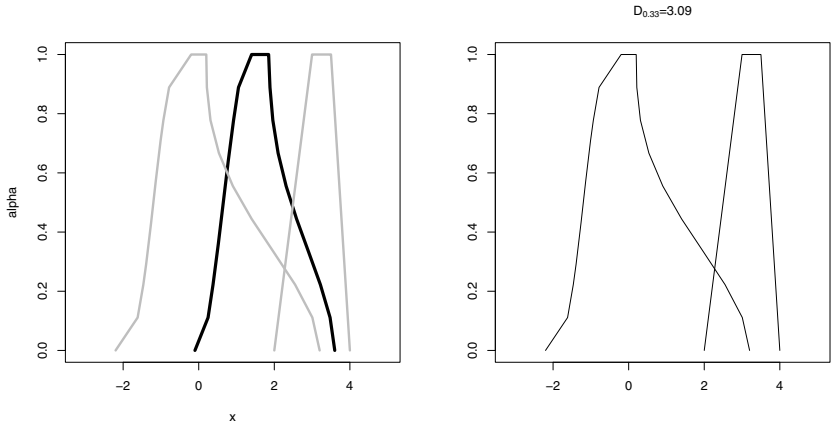


Fig. 2 The two fuzzy numbers (gray), their mean (black), and their Bertoluzza distance $d_{1/3}(X, Y)$ in Example 1

approach described in [4] which essentially imitates the Fourier series representation of every element in a separable Hilbert space w.r.t. an orthonormal basis. More precisely, given an input dataframe `V` in the correct format (which will be the expectation of the simulated FRV) first `decomposer(V)` is called and the resulting dataframe contains the 'coordinates' of `V` with respect to a certain 'basis' (see [4]). These 'coordinates' are perturbated stochastically in order to generate a new polygonal fuzzy number. The distributions used for these perturbations can be selected in the call of the function, however, in the current version 0.3 only a few choices are possible: (1) The perturbation of the centre of the 1-cut `pertV` has to be of the form $\mathcal{N}(0, \sigma)$ or $\mathcal{U}(-a, a)$ with $\sigma, a > 0$. (2) The perturbation of the left part of the fuzzy set `pertL` has to be χ_1, $Exp(1)$ or $\ln \mathcal{N}(a, b)$ with expectation one. (3) The perturbation of the right part of the fuzzy set `pertR` has to be of the same form as that for the left part. A precise description of the procedure can be found in [4].

The code in the following example generates samples $X_1, X_2, \cdots, X_n$ ($n = 10, 100, 1.000, 10.000$) of a fuzzy random variable X with given expectation `V=translator(XX[[3]],101)` by using the default perturbations in the `generator` function, calculates $d_{1/3}(V, \overline{X}_n)$ and plots the first ten sample elements and $\overline{A}_n$. The resulting plot is depicted in Figure 3.

Example 2:

```
nf <- layout(matrix(c(1,2,3,4), 2, 2, byrow=TRUE), respect=TRUE)
V<-translator(XX[[3]],101)
ss<-c(10,100,1000,10000)
for (j in 1:4){
 YY<-vector("list",length=ss[j])
   for(i in 1:ss[j]){
    YY[[i]]<-generator(V,,,)
```

```
    }
   M<-Mmean(YY)
   dis<-round(bertoluzza(M,V,1/3,0),4)
   plot(M,type="l",xlim=c(-4,4),main=paste("sample_size: ", ss[j],", D=",
    dis,sep=""), lwd=2, cex.main=1)
   lines(V,type="l",col="red",lwd=2)
    for (k in 1:10){
     lines (YY[[k]],type="l",col="gray")
    }
}
```

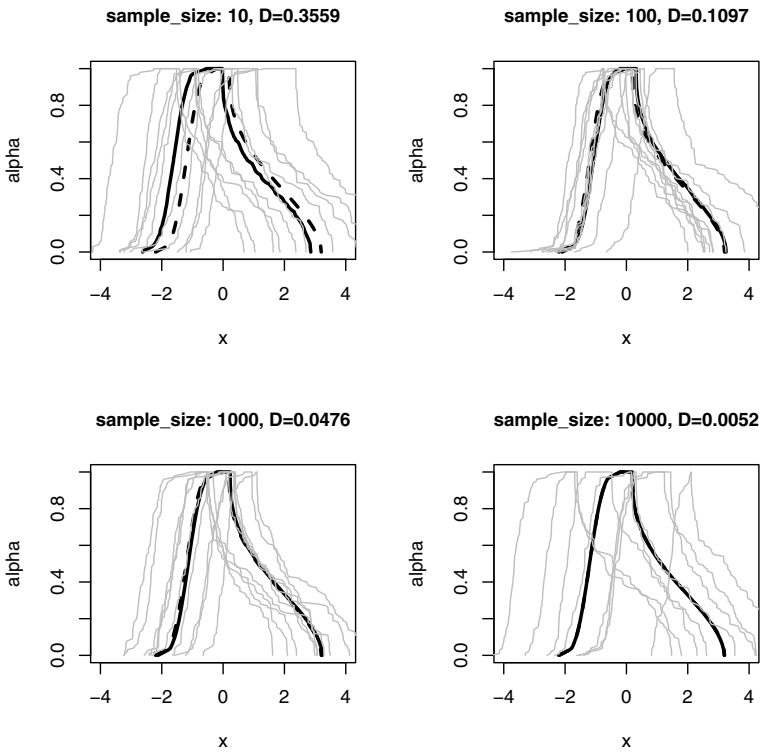


Fig. 3 Output of Example 2; the black line depicts the theoretical expectation, the dashed black line the sample mean

Figure 4 depicts the corresponding results for the case of choosing the following perturbations in Example 2:

```
pertV<-list(dist="unif",par=c(-2,2))
pertL<-list(dist="lnorm",par=c(-2,2))
```

Given a sample $\mathfrak{X} = (A_1, A_2 \ldots, A_n)$ of polygonal fuzzy numbers like `YY`, in Example 2 one can apply `DSfrequency` and `DShistogram` to calculate the

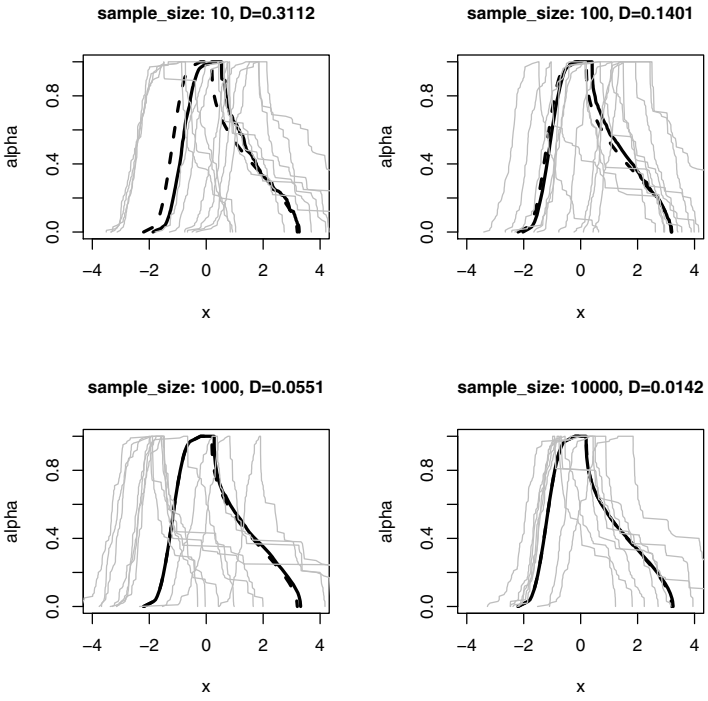


Fig. 4 Output of the modified version of Example 2

(Dempster-Shafer) frequency $H_n(I)$ of an interval I and a (Dempster-Shafer) histogram of $\mathfrak{A}$ respectively. The code in the following example generates data like in Example 2 and then calculates the frequency of $[1, 3]$ (201 different α-levels) - first with sample size $n = 3$ and afterwards with sample size $n = 50$. Figure 5 depicts the results (after gathering the individual plots produced by the functions in a joint one).

Example 3:

```
SS<-vector("list",length=3)
for (j in 1:3){
 SS[[j]]<-generator(V,)
 }
A<-DSfrequency(SS,c(1,3),1,201)

SS<-vector("list",length=50)
for (j in 1:50){
 SS[[j]]<-generator(V,)
 }
A<-DSfrequency(SS,c(1,3),1,201)
```

The code in Example 4 generates a sample of $n = 2.000$ fuzzy numbers and calculates the Dempster-Shafer histogram for the interval $[-4, 4]$ with

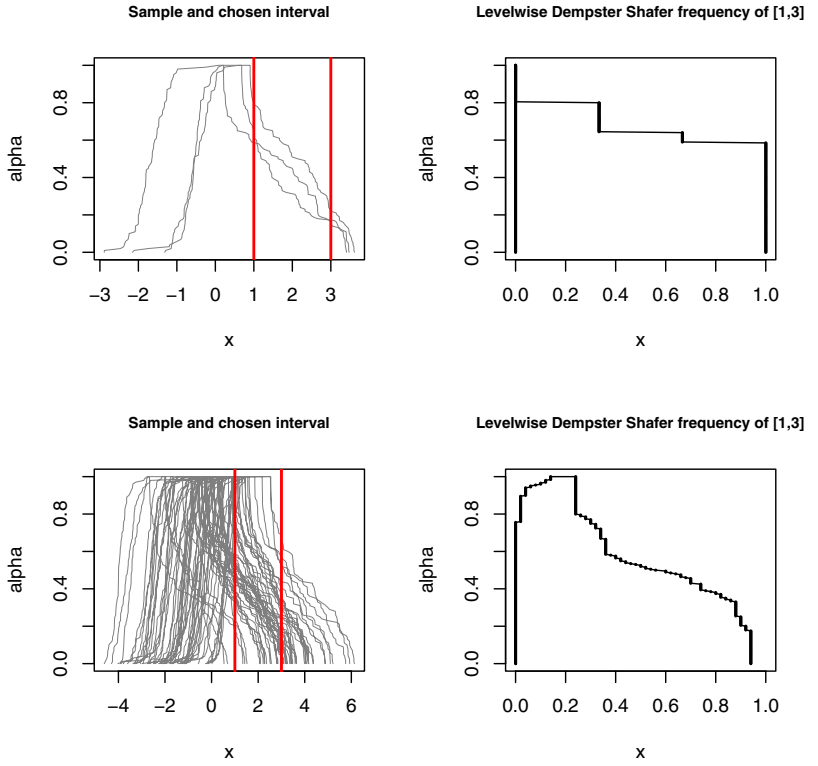


Fig. 5 Results of the code in Example 3

20 partition elements. By default `DShistogram` produces two plots, a $3d$-plot of the histogram as well as an image plot with the same colour scale, see Figure 6.

Example 4:

```
V<-translator(XX[[3]],51)
V$x<-V$x/10
SS<-vector("list",length=2000)
for (j in 1:2000){
 SS[[j]]<-generator(V,)
 }
A<-DShistogram(SS,c(-4,4),npart=20,nl=51,pdf=TRUE)
```

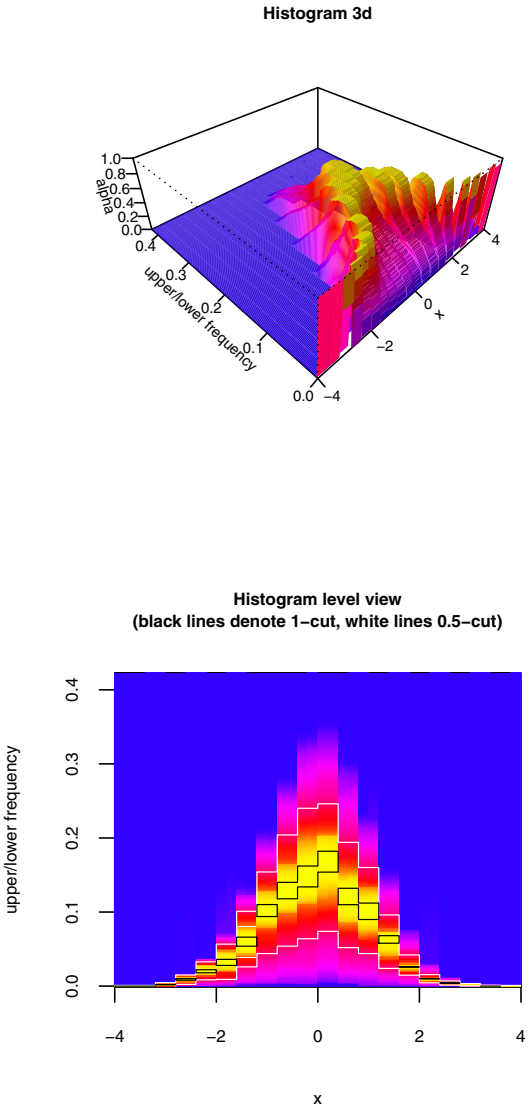


Fig. 6 Results of the code in Example 4

3.3 Further Functions in SAFD

As mentioned in the introduction SAFD also contains functions for bootstrap tests for the equality of means of fuzzy random variables and a function for calculating quantiles of samples of fuzzy numbers. Concerning these two functions we refer to [6, 9].

4 Future Work

In its next version the SAFD package will also contain functions for calculating the empirical lower and upper α-level distribution functions (defined analogously to the frequencies in (6)) and some robust version of basic functions. Furthermore the `generator` function will allow for more flexibility. Apart from that we plan to improve the speed both of the implemented bootstrap tests for the equality of means and of the `DSfrequency` and `DShistogram` functions, and allow more flexible plotting.

Acknowledgements This research has been partially supported by / benefited from the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02, and the COST Action IC0702. This financial support is gratefully acknowledged.

References

1. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers. *Mathware & Soft Comput.* 2:71–84
2. Colubi A (2009) Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data. *Fuzzy Sets and Systems* 160(3):344–356
3. González-Rodríguez G, Blanco A, Colubi A, Lubiano MA (2009) Estimation of a simple linear regression model for fuzzy random variables. *Fuzzy Sets and Systems* 160(3):357–370
4. González-Rodríguez G, Colubi A, Trutschnig W (2009) Simulation of fuzzy random variables. *Information Sciences* 179(5):642–653
5. Hesketh B, Hesketh T, Hansen J-I, Goranson D (1995) Use of fuzzy variables in developing new scales from the strong interest inventory. *J. Counseling Psychology* 42:85–99.
6. Lubiano MA, Trutschnig W (2010) ANOVA for Fuzzy Random Variables Using the R-package SAFD. In: Borgelt C *et al.* (eds) *Combining Soft Computing and Statistical Methods in Data Analysis*, 449–456. Springer-Verlag, Berlin Heidelberg
7. Möller B, Beer M (2004) *Fuzzy Randomness — Uncertainty in Civil Engineering and Computational Mechanics*. Springer, Berlin
8. Puri ML, Ralescu DA (1986) Fuzzy random variables. *J. Math. Anal. Appl.* 114:409–422
9. Sinova B, Gil MA, Colubi A, Van Aelst S (2012) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets and Systems* (to appear)
10. Trutschnig W (2008) A strong consistency result for fuzzy relative frequencies interpreted as estimator for the fuzzy-valued probability. *Fuzzy Sets and Systems* 159(3):259–269
11. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Information Sciences* 179(23):3964–3972
12. Viertl R, Hareter D (2006) *Beschreibung und Analyse unscharfer Information: Statistische Methoden für unscharfe Daten*. Springer, Wien New York
13. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Part 1. *Inform. Sci.* 8:199–249; Part 2. *Inform. Sci.* 8:301–353; Part 3. *Inform. Sci.* 9:43–80

Statistical Reasoning with Set-Valued Information: Ontic vs. Epistemic Views

Didier Dubois[1]

Abstract Sets, hence fuzzy sets, may have a conjunctive or a disjunctive reading. In the conjunctive reading a (fuzzy) set represents an object of interest for which a (gradual rather than Boolean) composite description makes sense. In contrast disjunctive (fuzzy) sets refer to the use of sets as a representation of incomplete knowledge. They do not model objects or quantities, but partial information about an underlying object or a precise quantity. In this case the fuzzy set captures uncertainty, and its membership function is a possibility distribution. We call epistemic such fuzzy sets, since they represent states of incomplete knowledge. Distinguishing between ontic and epistemic fuzzy sets is important in information-processing tasks because there is a risk of misusing basic notions and tools, such as distance between fuzzy sets, variance of a fuzzy random variable, fuzzy regression, etc. We discuss several examples where the ontic and epistemic points of view yield different approaches to these concepts.

1 Introduction

Traditional views of engineering sciences aim at building a mathematical model of a real phenomenon, via a data set containing observations of the concerned phenomenon. This mathematical model is approximate in the sense that it is an imperfect copy of the reality it intends to account for, but it is often precise, namely it typically takes the form of a real-valued function that represents, for instance, the evolution of a quantity over time. Approaches vary according to the class of functions used. The oldest and most common class is the one of linear functions, but a lot of works dealing with non-linear models have appeared, for instance and prominently, using neural networks

[1] IRIT, CNRS and Université de Toulouse, France, dubois@irit.fr

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 119–136.
springerlink.com

and fuzzy systems. These two techniques for constructing precise models have been merged to some extent due to the great similarity between the mathematical account of fuzzy rules and neurons, and their possible synergy due to the joint use of linguistic interpretability of fuzzy rules and learning capabilities of neural nets. While innovative with respect to older modeling techniques, these methods remain in the traditional school of producing a simplified and imperfect substitute of reality as observed via precise data.

Besides, there also exists a strong tradition of accounting for the non-deterministic aspect of many real phenomena subject to randomness in repeated experiments, including the noisy environment of measurement processes. Stochastic models enable to capture the general trends of populations of observed events through the use of probability distributions having a frequentist flavor. The probability measure attached to a quantity then reflects its variability through observed statistical data. Again in this approach, a stochastic model is a precise description of variability in physical phenomena.

More recently, with the emergence of Artificial Intelligence, but also in connection with more traditional human-centered research areas like Economics, Decision Analysis and Cognitive Psychology, the concern of reasoning about knowledge has emerged as a major paradigm [29]. While this topic has been mainly developed in the framework of classical or modal logic, due to the long philosophical tradition in this area, it has strongly affected the development of new uncertainty theories [20], and has led to a critique of probability theory as a unique framework for the representation of variability and belief. These developments question traditional views of modeling as representing reality independently of perception. They suggest a different approach that should also account for the cognitive limitations of our observations of reality. In other words, one might think of developing the epistemic approach to modeling. We call *ontic model* a precise representation of reality (however inaccurate it may be), and *epistemic model* a mathematical representation both of reality and the knowledge of reality, that explicitly accounts for the limited precision of our measurement capabilities. Typically, while the output of an ontic model is precise (but possibly wrong), an epistemic model delivers an imprecise output (hopefully consistent with the reality it accounts for). An epistemic model should of course be as precise as possible, given the available incomplete information, but it should also be as plausible as possible, avoiding unsupported arbitrary precision.

This paper discusses epistemic modeling in the context of set-based representations, and the mixing of variability and incomplete knowledge as present in recent works in fuzzy set-valued statistics.

2 Ontic vs. Epistemic Sets

A set S defined in extension, is often denoted by listing its elements, say, in the finite case $\{s_1, s_2, \ldots, s_n\}$. As pointed out in a recent paper [21] this representation, when it must be used in applications, is ambiguous. In some cases, a set represents a real complex lumped entity. It is then a conjunction of its elements. It is a precisely described entity made of subparts. For instance, a region in a digital image is a conjunction of adjacent pixels; a time interval spanned by an activity is the collection of instants where this activity takes place. In other cases, sets are mental constructions that represent incomplete information about an object or a quantity. In this case, a set is used as a disjunction of possible items, or of values of this underlying quantity, one of which is the right one. For instance I may only have a rough idea of the birth date of the president of some country, and provide an interval as containing this birth date. Such an interval is the disjunction of mutually exclusive elements. It is clear that the interval itself is subjective (it is my knowledge), has no intrinsic existence, even if it refers to a real fact. The use of sets representing imprecise values can be found for instance in interval analysis [39]. Another example is the set of models of a propositional knowledge base: only one of them reflects the real situation. Moreover this set is likely to change by acquiring more information.

Sets representing collections C of elements forming composite objects will be called *conjunctive*; sets E representing incomplete information states will be called *disjunctive*. A conjunctive set is the precise representation of an objective entity (philosophically it is a *de re* notion), while a disjunctive set only represents incomplete information (it is *de dicto*). We also shall speak of *ontic* sets, versus *epistemic* sets, in analogy with ontic vs. epistemic actions in cognitive robotics [30]. An ontic set C is the value of a set-valued variable X (and we can write $X = C$). An epistemic set E contains the ill-known actual value of a point-valued quantity x and we can write $x \in E$. A disjunctive set E represents the epistemic state of an agent, hence does not exist per se. In fact, when reasoning about an epistemic set it is better to handle a pair (x, E) made of a quantity and the available knowledge about it.

A value s inside a disjunctive set E is a possible candidate value for x, while elements outside E are considered impossible. Its characteristic function can be interpreted as a possibility distribution [56]. This distinction between conjunctive and disjunctive sets was made by Zadeh himself [57] distinguishing between set-valued attributes (like the set of sisters of some person) from ill-known single-valued attributes (like the unknown single sister of some person). This issue has been extensively discussed by Yager [53] and Dubois and Prade [17] for the study of incomplete conjunctive information (whose representation requires a disjunctive set of conjunctive sets).

An epistemic set (x, E) does not necessarily accounts for an ill-known deterministic value. An ill-known quantity may be deterministic or stochastic. For instance, the birth date of a specific individual is not a random variable

even if it can be ill-known. On the other hand the daily rainfall in a specific place is a stochastic variable, since it can be modelled by a probability distribution. An epistemic set then captures in a rough way information about a population via observations. For instance, there is a sample space Ω, and x can be a random variable taking values on S, but the probability distribution on Ω is unknown. All that is known is that $x(\omega) \in E$, that is $P_x(E) = 1$ where P_x is the probability measure of x. In that case, E represents the family $\mathcal{P}_E$ of objective probability measures on Ω such that $P(\{\omega : x(\omega) \in E\}) = 1$, one of which being the proper representation of the random phenomenon. In this case, the object to which E refers is not a precise value of x, but a probability measure P_x describing the variability of x.

Note that in the probabilistic literature, an epistemic set is more often than not modelled by a probability distribution. In the early 19th century, Laplace proposed to use a uniform probability on E, based on the insufficient reason principle, according to which what is equipossible must be equiprobable. This is a default choice in $\mathcal{P}_E$ that coincides with the probability distribution having maximal entropy. However, this approach makes sense if x is a random variable. In case x is an ill-known deterministic value, Bayesians [35] propose to use a subjective probability P_x^b in place of set E. In that case, where the occurrence of x is not a matter of repetitions, $P_x^b(A)$ is the price of a lottery ticket chosen by an agent who agrees to earn \$1 if A turns out to be true, in an exchangeable bet scenario where the bookmaker exchanges roles with the buyer if the proposed price is found unfair. It forces the agent to propose prices $p^b(s)$ that sum exactly to 1 over E. Then $P_x^b(A)$ measures the degree of belief of the (non-repeatable) event $x \in E$, and this degree is agent-dependent.

However clever it may be, this view is debatable (see [20] for a summary of critiques). Especially, this representation is unstable: if P_x^b is uniform on E, then $P_{f(x)}^b$ may fail to be so if E is finite and the image $f(E)$ does not contain the same number of elements as E, or if E is an interval and f is not a linear transformation. Moreover, the use of unique probability distributions to represent belief is challenged by experimental results (like Ellsberg paradox [4]), which show that individuals do not make decisions based on expected utility in front of partial ignorance.

3 Random Sets vs. Ill-known Random Variables

As opposed to the case of an epistemic set representing an ill-known probability distribution, another situation is when the probability space (Ω, P) is available[1], but each realisation of the random variable is represented as a set. This case covers two situations:

[1] In this paper, we assume Ω is finite to avoid mathematical difficulties.

1. **Random conjunctive sets**: The random variable $X(\omega)$ is multi-valued and takes values on the power set of a set S. For instance, S is a set of spoken languages, and $X(\omega)$ is the set of languages spoken by an individual ω. Or $X(\omega)$ is an ill-known area of interest in some spatial domain, and ω is the outcome of an experiment to locate it. Then a probability distribution p_X is obtained over 2^S, such that $p_X(C) = P(X = C)$. It is known in the literature as a random set (Kendall [31], Matheron [38]). In our terminology this is a random conjunctive (or ontic) set.
2. **Ill-known random variables**: The random variable $x(\omega)$ takes values on S but its realisations are incompletely observed. It means that $\forall \omega \in \Omega$, all that is known is that $x(\omega) \in E = X(\omega)$ where X is a multiple-valued mapping $\Omega \to 2^S$ representing the disjunctive set of mappings (called selections) $\{x : \Omega \to S, \forall \omega, x(\omega) \in X(\omega)\} = \{x \in X\}$ for short. In other words the triple (Ω, P, X) is an epistemic model of the random variable x. This is the approach of Dempster [11] to imprecise probabilities. He uses this setting to account for a parametric probabilistic model P_θ on a set U of observables, where $\theta \in \Theta$ is an ill-known parameter but the probability distribution of a function $\phi(u, \theta) \in \Omega$ is known. Then $S = \Theta \times U$ and $X(\omega) = \{(\theta, u), \exists \theta, \phi(u, \theta) = \omega\}$. It is clear that for each ω, $X(\omega)$ is an epistemic set restricting, for each observation u the actual (deterministic) value θ.

Shafer [46] has proposed a non-statistical view of the epistemic random set setting, based on a subjective probability m over 2^S, formally identical to p_X. In this setting called the theory of evidence, $m(E)$ represents the subjective probability that all that is known of a deterministic quantity x is of the form $x \in E$. This is the case when an unreliable witness testifies that $x \in E$ and p is the degree of confidence of the receiver agent in the validity of the testimony. Then with probability $m(E) = p$, $x \in E$ is a reliable information. It means that the testimony is useless with probability $m(S) = 1 - p$ assigned to the empty information S. This view of probability was popular until the end of the 18th century (see [41] for details and a general model of unreliable witness). More generally the witness can be replaced by a measurement device or a message-passing entity with state space U, such that if the device is in state u then the available information is of the form $x \in E(u) \subseteq S$, and $p(u)$ is the subjective probability that the device is in state u [47].

The above discussions lay bare the difference between random conjunctive and disjunctive sets, even if they share the same mathematical model. In the first case one may compute precise probabilities that a set-valued variable X takes value in a family $\mathcal{A}$ of subsets:

$$P_X(\mathcal{A}) = \sum_{X(\omega) \in \mathcal{A}} p(\omega) = \sum_{C \in \mathcal{A}} p_X(C). \tag{1}$$

For instance, in the language example, and $S = \{\text{English, French, Spanish}\}$, one may compute the probability that someone speaks English by summing

the proportions of people in Ω that respectively speak English only, English and French, English and Spanish, and the three languages.

In the second scenario, the random set $X(\omega)$ represents knowledge about a point-valued random variable $x(\omega)$. For instance, suppose S is an ordered height scale, $x(\omega)$ represents the height of individual ω and $X(\omega) = [a, b] \subseteq S$ is an imprecise measurement of $x(\omega)$. Here one can compute a probability range containing the probability $P_x(A) = \sum_{x(\omega)\in A} p(\omega)$ that the height of individuals in Ω lies in A, namely lower and upper probabilities proposed by Dempster [11]:

$$\underline{P}_X(A) = \sum_{X(\omega)\subseteq A} p(\omega) = \sum_{E\subseteq A} p_X(E); \tag{2}$$

$$\overline{P}_X(A) = \sum_{X(\omega)\cap A\neq\emptyset} p(\omega) = \sum_{E\cap A\neq\emptyset} p_X(E) \tag{3}$$

such that $\underline{P}_X(A) = 1 - \overline{P}_X(\bar{A})$, where $\bar{A}$ is the complement of A. Note that the set of probabilities $\mathcal{P}_X$ on S induced by this process is finite: since Ω and S are finite, the number of selections $x \in X$ is finite too. In particular, $\mathcal{P}_X$ is not convex. Its convex hull is $\tilde{\mathcal{P}}_X = \{P_S; \forall A \in S, P_S(A) \geq \underline{P}_X(A)\}$. It is well-known that probability measures in this convex set are of the form

$$P_S(A) = \sum_{E\subseteq S} p_X(E) P_E(A)$$

where P_E, a probability measure such that $P_E(E) = 1$, defines a sharing strategy of probability weight $p_X(E)$ among elements of E. As explained by Couso and Dubois [7], it corresponds to a scenario where when $\omega \in \Omega$ occurs, $x(\omega)$ is tainted with variability (due to the measurement device) that can be described by a conditional probability $P(\cdot|\omega)$ on S. Hence the probability $P_x(A)$ is now of the form:

$$P_x(A) = \sum_{\omega\in\Omega} P(A|\omega)P(\omega).$$

However, all we know is that $P(X(\omega)|\omega) = 1$ for some maximally specific epistemic subset $X(\omega)$. This is clearly a third (epistemic) view of the random set X. It is easy to see that the choice of $\mathcal{P}_X$ vs. its convex hull is immaterial in the computation of upper and lower probabilities, so that

$$\underline{P}_X(A) = \inf\{\sum_{\omega\in\Omega} P(A|\omega)P(\omega) : P(X(\omega)|\omega) = 1, \forall\omega \in \Omega\} \tag{4}$$

$$= \inf\{\sum_{E\subseteq S} p_X(E)P_E(A) : P_E(E) = 1\}. \tag{5}$$

where $P_E(A) = P(A|\omega)$ if $E = X(\omega)$.

In the evidence theory setting, Dempster upper and lower probabilities of an event are directly interpreted as degrees of belief $Bel(A) = \underline{P}_X(A)$ and plausibility $Pl(A) = \overline{P}_X(A)$, without reference to an ill-known probability on S (since the information is not frequentist here). This is the view of Smets [49]. The mathematical similarity between belief functions and random sets was quite early pointed out by Nguyen [40]. But they gave rise to quite distinct streams of literature that tend to ignore or misunderstand each other.

4 When the Meaning of the Model Affects Results

The reader may consider that the three above interpretations of random sets are just a philosophical issue, but do not impact on computations that can be carried out with this model. For instance the mean interval of a random interval has the same definition (interval arithmetics or Aumann integral) independently of the approach. However this is not true for other concepts. Two examples are given: conditioning and variance.

4.1 Conditioning Random Sets

Given a random set in the form of a probability distribution on the power set S, and an event $A \subset S$, the proper method for conditioning the random set on A depends on the adopted scenario.

Conditioning a conjunctive random set In this case the problem comes down to restricting the set-valued realisations $X(\omega)$ so as to account for the information that the set-valued outcome lies inside A. Then the obtained conditional random set is defined by means of the standard Bayes rule in the form of its weight distribution $p_X(\cdot|A)$ such that:

$$p_X(C|A) = \begin{cases} \frac{p_X(C)}{\sum_{B \subseteq A} p_X(B)} \text{ if } C \subseteq A; \\ 0 \text{ otherwise.} \end{cases} \tag{6}$$

Conditioning an ill-known random variable Suppose the epistemic random set $X(\omega)$ relies on a population Ω, and is represented by the convex set of probabilities $\tilde{\mathcal{P}}_X$ on S, one of which is the proper frequentist distribution of the underlying random variable x. Suppose we study a case for which all we know is that $x \in A$, and the problem is to predict the value of x. Each probability $p_X(E)$ should be altered in order to restrict to the subset $\Omega_A = \{\omega : x(\omega) \in A\}$ of population Ω. However, because $x(\omega)$ is only known to lie in $X(\omega)$, the set Ω_A is itself ill-known. There are three situations:

1. Either $A \cap E = \emptyset$: then $\Omega_A \cap \{\omega : X(\omega) = E\} = \emptyset$ and we can drop $p_X(E)$.
2. Or $E \subseteq A$ and then $\{\omega : X(\omega) = E\} \subseteq \Omega_A$ and $p_X(E)$ should remain assigned to E;
3. Or E overlaps both A and its complement: then let $\alpha_A(E)$ be the proportion of the population for which all we know is $x(\omega) \in E$ and that lies inside Ω_A. The weight $\alpha_A(E)p_X(E)$ should be assigned to $E \cap A$.

One may then define the conditional probability distribution over 2^S as follows:

$$p_X^{\alpha_A}(B|A) = \begin{cases} \frac{\sum_{B=E\cap A} \alpha_A(E)p_X(E)}{\sum_{E\cap A\neq\emptyset} \alpha_A(E)p_X(E)} \text{ if } B \subseteq A; \\ 0 \text{ otherwise.} \end{cases} \tag{7}$$

This mass assignment leads to computing lower and upper probabilities $\underline{P}^{\alpha_A}(\cdot|A)$ and $\overline{P}^{\alpha_A}(\cdot|A)$ when the vector of weights α_A is fixed. But this proportion $\alpha_A(E)$ is unknown in the third situation, while it is respectively 0 and 1 in the previous ones. Varying this unknown vector leads to upper and lower conditional probabilities as follows:

$$\overline{P}_X(B|A) = \sup_{\alpha_A} \overline{P}^{\alpha_A}(B|A); \qquad \underline{P}_X(B|A) = \inf_{\alpha_A} \underline{P}^{\alpha_A}(B|A). \tag{8}$$

and likewise for the lower conditional probability. In fact, it has been proved that these bounds can be obtained by applying Bayesian conditioning to all probabilities in $\tilde{\mathcal{P}}_X$ with $P_x(A) > 0$ and that they take an attractive closed form [9, 23]:

$$\overline{P}_X(B|A) = \sup\{P_x(B|A) : P_x \in \tilde{\mathcal{P}}_X\} \quad = \frac{\overline{P}_X(B \cap A)}{\overline{P}_X(B \cap A) + \underline{P}_X(\bar{B} \cap A)}, \tag{9}$$

$$\underline{P}_X(B|A) = \inf\{P_x(B|A) : P_x \in \tilde{\mathcal{P}}_X\} \quad = \frac{\underline{P}_X(B \cap A)}{\underline{P}_X(B \cap A) + \overline{P}_X(\bar{B} \cap A)}, \tag{10}$$

where $\underline{P}_X(B|A) = 1 - \overline{P}_X(\bar{B}|A)$ and $\bar{B}$ is the complement of B.

Conditioning a belief function In this case, there is no longer any population, and the probability distribution $m = p_X$ on 2^S represents subjective knowledge about a deterministic value x. Conditioning on A means that we come to hear that the actual value of x lies in A for sure. Then we perform an information fusion process (a special case of Dempster rule of combination [11]). It yields yet another type of conditioning, called Dempster conditioning, that systematically transfers masses $m(E)$ to $E \cap A$ when not empty, eliminates $m(E)$ otherwise, then normalises the conditional mass function, dividing by $\sum_{E\cap A\neq\emptyset} m(E) = Pl(A)$. It leads to the conditioning rule

$$Pl(B|A) = \frac{Pl(A \cap B)}{Pl(A)} = \frac{\overline{P}_X(A \cap B)}{\overline{P}_X(A)}, \tag{11}$$

and $Bel(B|A) = 1 - Pl(\bar{B}|A)$. Note that it comes down to the previous conditioning rule (7) with $\alpha_A(E) = 1$ if $E \cap A \neq \emptyset$, and 0 otherwise (an optimistic assignment, justified by the claim that A contains the actual value of x). Interestingly the conditioning rule for conjunctive random sets comes down to the previous conditioning rule (7) with $\alpha_A(E) = 1$ if $E \subseteq A$, and 0 otherwise, that could, in the belief function terminology, be written as $Bel(B|A) = \frac{Bel(A \cap B)}{Bel(A)}$. It is known as the geometric rule of conditioning. Such a pessimistic weight reassignment can hardly be justified for disjunctive random sets.

4.2 Empirical Variance for Random Interval Data

Interval data sets provide a more concrete view of a random set. Again the distinction between the case where such intervals represent precise actual objects and when they express incomplete knowledge of precise ill-observed point values is crucial in computing a statistical parameter such as variance [7]. Consider a data set consisting of a bunch of intervals $\mathbb{D} = \{I_i = [\underline{a}_i, \overline{a}_i], i = 1, ...n\}$. The main question is: are we interested by the joint variation of the size and location of the intervals ? or are we interested in the variation of the underlying precise quantity as imperfectly accounted for by the variation of the interval data?

1. **Ontic interval data**: In this case we consider intervals are precise lumped entities. For instance, one may imagine the interval data set to contain sections of a piece of land according to coordinate x in the plane: $I_i = Y(x_i)$ for a multimapping Y, where $Y(x_i)$ is the extent of the piece of land at abscissa x_i, along coordinate y. The ontic view suggests the use of a scalar variance:

$$ScalVar(\mathbb{D}) = \frac{\sum_{i=1,...,n} d(M, I_i)^2}{n}, \tag{12}$$

   where $M = [\sum_{i=1}^n \underline{a}_i/n, \sum_{i=1}^n \overline{a}_i/n]$ is the interval mean value, and d is a scalar distance between intervals (e.g. Euclidean distance between pairs of values representing the endpoints of the intervals, but more refined distances have been proposed [2]). $ScalVar(\mathbb{D})$ measures the variability of the intervals in $\mathbb{D}$, both in terms of location and width and evaluates the spatial regularity of the piece of land, varying coordinate x. This variance is 0 for a rectangular piece of land parallel to the coordinate axes.
2. **Epistemic interval data**: Under the epistemic view, each interval I_i stands for an ill-known precise value x_i that is the result of measuring a deterministic value x several times. Here, the measurement process is subject to randomness and is imprecise. Then we are more interested by sensitivity analysis describing what we know about the variance we would

have computed, had the data been precise. Then, we should compute the interval

$$EVar_1(\mathbb{D}) = \{var(\{x_i, i = 1, ...n\}) : x_i \in I_i, \forall i\}. \quad (13)$$

Computing this interval is a non-trivial task [25, 48]

3. **Epistemic interval random data** Alternatively one may consider that the quantity x that we wish to describe is intrinsically random. Each measurement process is an information source providing incomplete information on the variability of x. Then each interval I_i can be viewed as containing the support $SUPP(P_i)$ of an ill-known probability distribution P_i: then we get a wider variance interval than previously. It is defined by

$$EVar_2(\mathbb{D}) = \{var(\sum_{i=1}^{n} P_i/n) : SUPP(P_i) \subseteq I_i, \forall i = 1, ...n\} \quad (14)$$

and it is easy to see that $EVar_1(\mathbb{D}) \subset EVar_2(\mathbb{D})$.

In the extreme case of a single epistemic interval $(x, [a, b])$, if x is a deterministic ill-known quantity, it has a unique true value. Then $EVar_1([a, b]) = var(x) = 0$ (since even if ill-known, x is not supposed to vary: the set of variances of a bunch of Dirac functions is $\{0\}$). In the second case, x is tainted with variability, $var(x)$ is ill-known and lies in the interval $EVar_2([a, b]) = [0, v^*]$ where $v^* = \sup\{var(x), SUPP(P_x) \subseteq [a, b]\} = (b - a)^2$. The distinction between deterministic and stochastic variables known via intervals thus has important impact on the computation of dispersion indices, like variance.

Note that in the epistemic view, the scalar distance between intervals can be useful. It is then a kind of informational distance between pieces of knowledge, whose role can be similar to relative entropy for probability distributions. Namely one may use it in revision processes, for instance. Moreover one may be interested by the scalar variance of the imprecision of the intervals, or by an estimate of the actual variance of the underlying quantity, by computing the variance of say the mid-points of the intervals. Recently suggested scalar variances [44] between intervals come down a mixture of such a scalar variability estimation and the variance of imprecision.

5 Different Interpretations of a Fuzzy Set

A fuzzy set on a universe S is mathematically modelled by a mapping from S to a totally ordered set L that is usually the unit interval. As highlighted by Dubois and Prade [19], a membership function is an abstract object that needs to be interpreted in practical settings in order to be used meaningfully. They proposed three interpretations of membership grades in terms of degrees of similarity, of plausibility and preference. An early and important use of fuzzy sets, proposed by Zadeh [55] is the representation of symbolic categories

on numerical universes. A linguistic variable is a variable that takes values on a set of linguistic terms modelled by fuzzy sets of the real line. In this case, degrees of membership express similarity or distance to prototypical values covered by a term.

As already acknowledged a long time ago, fuzzy sets, like sets, may have a conjunctive or a disjunctive reading [57, 53, 17]. In the conjunctive reading, ontic fuzzy sets represent objects originally construed as sets but for which a fuzzy representation is more expressive due to gradual boundaries. Degrees of membership evaluate to what extent components participate to the global entity. For instance, this is the case when modeling linguistic labels by convex fuzzy sets on a measurable scale, like *tall, medium-sized, short* achieving a fuzzy partition of the human height scale. In this case, the fuzzy sets have a conjunctive reading because they are understood as the set of heights compatible with a given label. Other examples of ontic fuzzy sets are non-Boolean classes stemming from a clustering process, fuzzy constraints representing preference, a fuzzy region in an image, a fuzzy rating profile according to various attributes. As a concrete example, consider the fuzzy set of languages more or less well spoken by a person.

In contrast, Zadeh [56] also proposed to interpret membership functions as possibility distributions, paving the way to a representation of incomplete information along a line followed thirty years earlier by Shackle [45]. In that case, a degree of membership refers to the idea of plausibility. A possibility distribution, denoted by π is the membership function of a fuzzy set of mutually exclusive values in S. A possibility distribution is supposedly attached to an ill-known quantity x. Namely $\pi(s) > 0$ expresses that s is a possible value of x, all the more plausible as $\pi(s)$ is greater. In particular it is assumed that $\pi(s) = 1$ for some value s, which is then considered as normal, totally unsurprizing. A possibility distribution thus extends the set-valued representation of incomplete information to account for degrees of plausibility. It is well-known that a possibility distribution π induces a possibility measure Π on 2^S such that $\Pi(A) = \sup_{s \in A} \pi(s)$ for all events A and a necessity measure $N(A) = 1 - \Pi(\bar{A})$ [16].

Now, if the information about a quantity x is expressed by means of a fuzzy set, the above distinction between the deterministic and the stochastic case is again at work. If x is deterministic, then this information must be interpreted in terms of "confidence sets" as follows. Let $E_\alpha = \{s, \pi(x) \geq \alpha\}$ be the α-cut of π:

For each $\alpha \in [0, 1]$, $x \in E_\alpha$ with probability greater than or equal to $1 - \alpha$.

If an expert provides this kind information, the word "probability" refers to subjective probability. Following Walley [52], $1 - \alpha$ is the maximal price at which this expert would buy the gamble that wins \$1 if the real value of x actually lies in E_α (the minimal selling price for this gamble is \$1). Note that there is no "real probability distribution" underlying π, but Dirac functions as x is deterministic. The consonance of the family of sets E_α makes sense

if this is the opinion of a single expert who tends to be imprecise but self-consistent.

If x is stochastic then there are two possible ways of interpreting the possibility distribution π.

- Mathematically speaking, a possibility measure is a coherent upper probability [52], namely $\Pi(A) = \sup_{P \in \mathcal{P}_\pi} P(A)$ where $\mathcal{P}_\pi = \{P, \forall A, P(A) \leq \Pi(A)\}$. So, π encodes the set of probabilities $\mathcal{P}_\pi$ [18, 14]. This set is supposed to contain the real probability measure P_x that governs the variability of x. It is a set-based representation of a stochastic variable representing incomplete information about a frequentist probability. An expert providing distribution π claims that

  For each $\alpha \in [0, 1]$, the event $x \in E_\alpha$ has *objective* probability greater than or equal to $1 - \alpha$.

- Another option is to consider π as encoding a higher-order (subjective) possibility distribution on a set of objective probabilities. Namely, it can be understood as follows:

  For each $\alpha \in [0, 1]$, P_x has support in E_α with subjective probability greater than or equal to $1 - \alpha$.

  So the domain of π can be canonically extended to the set of probability measures on S as follows: $\pi(P) = \sup\{\alpha, P \text{ has support in } E_\alpha\}$. The possibility measure Π is a "second-order possibility" formally equivalent to those considered in [10]. It is so called, because it is a possibility distribution defined over a set of probability measures. The deterministic case is a special case of this framework, restricting probability measures to Dirac measures. It would be interesting to investigate the relationship between the set of probabilities $\mathcal{P}_\pi$ and the higher order possibility model.

The above setting does not make it clear where the objective probability distribution comes from, i.e. the underlying sample space. Moreover, it does not account for the measurement process of x. Namely, regardless of whether x is deterministic or stochastic, there may be a stochastic measurement process yielding with more or less accuracy information on the possible values of x. The setting of fuzzy random variables extends the above distinctions by taking the measurement process into account explicitly.

6 Various Notions of Random Fuzzy Sets

The history of fuzzy random variables is not simple as it was started by two separate groups with respectively epistemic and ontic views in mind. The first papers are those of Kwakernaak [33, 34] in the late seventies, clearly underlying an epistemic view of fuzzy sets, a line followed up by Kruse and Meyer

[32]. They view a fuzzy random variable as a (disjunctive) fuzzy set of classical random variables (those induced by selection functions compatible with the random fuzzy set). It represents what is known about the variability of the underlying ill-known random variable. These works can thus be viewed as extending the framework of Dempster's upper and lower probabilities based on the triple (Ω, P, X) to fuzzy set-valued mappings $\tilde{X}$, where $\tilde{X}(\omega)$ defines a possibility distribution restricting the possible values of $x(\omega)$. The degree of possibility that x is the random variable underlain by $(\Omega, P, \tilde{X})$ is

$$\pi(x) = \inf_{\omega \in \Omega} \mu_{\tilde{X}(\omega)}(x(\omega)) \tag{15}$$

For each level $\alpha \in (0, 1], \tilde{X}_\alpha(\omega) = \{s \in S : \mu_{\tilde{X}(\omega)}(s) \geq \alpha\}$ is a multiple valued mapping such that $(\Omega, P, \tilde{X}_\alpha)$ is an epistemic random set according to Dempster framework. Kruse and Meyer [32] clearly define the variance of a fuzzy random variable as a fuzzy set of positive reals induced by applying the extension principle to the variance formula. Likewise, the probability of an event becomes restricted by a fuzzy interval in the real line [1]. The evidence theory counterpart of this view deals with belief functions having fuzzy focal elements [54]. An alternative epistemic view of fuzzy random variables was more recently proposed in the spirit of Walley [52], in terms of a convex set of probabilites induced on S[8].

In contrast, the line initiated in the mid-1980's by Puri and Ralescu [43] is in agreement with conjunctive random set theory. A fuzzy random variable is then viewed as a random conjunctive fuzzy set, i.e. a classical random variable ranging in a set of (membership) functions. This line of research has been considerably extended so as to adapt classical statistical methods to functional data [5, 27]. The main issue is to define a space of functions equipped with a suitable metric structure [13, 51]. In this theory of random fuzzy sets, a scalar distance between fuzzy sets is instrumental when defining variance viewed as a mean squared deviation from the fuzzy mean value [28], in the spirit of Fréchet. A scalar variance can be established on this basis and it reflects the variability *of membership functions*. It makes sense if for instance, membership functions are models of linguistic terms and some "term variability" must be evaluated given a set of responses provided by a set of people in natural language. See [7] for an extensive comparison of the three views of fuzzy random variables.

The ontic view is advocated by Colubi *et al.* [6] in the statistical analysis of linguistic data. The authors argue that they are interested in the statistics of perceptions. One of their experiments deals with the visual perception of the length of a line segment expressed on fuzzy scale using a linguistic label among *very small, small, medium, large, very large*. The alleged goal is to predict the category that a person considers correct for the segment. The precise length of the segment exists but it is irrelevant for the classification goal. They agree that to predict the real length from the fuzzy perceptions requires a different approach.

The case of Likert scaling is more problematic. This is a method of ascribing quantitative values to qualitative data, to make them amenable to statistical analysis. For instance, an ordered set of linguistic labels referring to some abstract concept (like beauty) is encoded by successive integers. A typical scale might be *strongly agree, agree, not sure/undecided, disagree, strongly disagree*. Opinions are collected on such a scale and a mean figure for all the responses is computed at the end of the evaluation or survey. A number of authors have proposed to model such linguistic terms by means of a predefined fuzzy partition made of fuzzy intervals (trapezoids) on a real interval. In some other approaches the format of the fuzzy response can be any fuzzy interval. The idea is to cope with the arbitrariness of encoding qualitative value by precise numbers. In that case the result of an opinion poll is clearly a random fuzzy set.

However this kind of approach is not convincing from a measurement point of view[15]. First, it is not clear why the underlying real interval can be equipped with addition at all. It is rather an ordinal scale, and trapezoidal fuzzy sets then make no sense. Next, this continuous scale is totally fictitious and it is patent that the real data are the linguistic terms provided by people: there is no underlying real value behind such linguistic terms. If the response has a free format (whereby any fuzzy interval can do), one may again see this fuzzy response as being the evaluation in itself. The latter point would plea for an ontic view of the random fuzzy sets. However the arbitrariness of the numerical encoding casts doubts on the cogency of the sophisticated functional analysis framework needed to apply fuzzy random set methods. It may be that ordinal statistical methods devoted to finite qualitative scales would be more appropriate in this case.

7 Epistemic vs. Ontic Interval Data Processing

Consider a set of bidimensional interval data $\mathbb{D} = \{(x_i, Y_i = [\underline{y}_i, \overline{y}_i]), i = 1, ...n\}$ or its fuzzy counterpart (if the Y_i's become fuzzy sets). The issue of devising an extension of data processing methods to such a situation has been studied in many papers in the last 20 years or so. But it seems that the question how the reading of the set-valued data has impact on the chosen method is seldom discussed. Here we provide some hints on this issue, restricting ourselves to linear regression and some of its fuzzy extensions.

A first approach that is widely known is Diamond's fuzzy least squares method [12]. It is based on a scalar distance between set-valued data. The problem is to find a best fit interval model of the form $y = A^*x + B^*$, where intervals A^*, B^* minimize $\sum_{i=1}^{n} d(Ax_i + B, Y_i)^2$, typically a sum of squares of differences between upper and lower bounds of intervals. The fuzzy least squares regression is similar but it presupposes the $\tilde{Y}_i'$s are triangular fuzzy intervals $(y_i^m; y_i^-, y_i^+)$, with modal value y_i^m and support $[y_i^-, y_i^+]$.

Diamond proposes to work with a scalar distance of the form $d^2(\tilde{A}, \tilde{B}) = (a^m - b^m)^2 + (a^- - b^-)^2 + (a^+ - b^+)^2$ making the space of triangular fuzzy intervals complete. The problem is then to find a best fit fuzzy interval model $\tilde{Y} = \tilde{A}^* x + \tilde{B}^*$, where fuzzy triangular intervals $\tilde{A}^*, \tilde{B}^*$ minimize a squared error $\sum_{i=1}^n d^2(\tilde{A}x_i + \tilde{B}, \tilde{Y}_i)$. Some comments are in order:

- This approach treats fuzzy data as ontic entities.
- If the (fuzzy) interval data-set is epistemic, we get a linear description of the trend of the knowledge as x increases.
- This approach does not correspond to studying the impact of data uncertainty on the result of regression.

Many variants of this method, based on conjunctive fuzzy random sets and scalar distances exist (see [24] for a recent one) including extensions to fuzzy-valued inputs [26]. These approaches all adopt the ontic view.

Another classical approach was proposed by Tanaka *et al.* in the early 1980's (see [50] for an overview). One way of posing the interval regression problem is to find a set-valued function $Y(x)$ (generally again of the form of an interval-valued linear function $Y(x) = Ax + B$) with maximal informative content such that $Y_i \subset Y(x_i), i = 1, ...n$. Some comments help situate this method:

- It does not presuppose an ontic or epistemic reading of the data. If data are ontic, the result models an interval-valued phenomenon. If epistemic, it tries to cover both the evolution of the variable y and the evolution of knowledge of this phenomenon.
- It does not clearly extend the basic concepts of linear regression.

Both approaches rely on the interval extension of a linear model $y(x) = ax+b$. But, in the epistemic reading, this choice imposes unnatural constraints on the relation between the epistemic output $Y(x)$ and the objective input x (e.g., $Y(x)$ becomes wider as x increases). The fact that the real phenomenon is affine does not imply that the knowledge about it in each point should be also of the form $Y(x) = Ax + B$. In an ontic reading, one may wish to interpolate the interval data more closely (see Boukezzoula *et al.* [3] for improvements of Tanaka's methods that cope with such defects).

Another view of interval regression, that has a clear epistemic flavor uses possibility theory to define a kind of quantile regression. Even when applied to precise data sets it gives an epistemic interval-valued representation of objective data, likely to contain the actual model [42]. The idea is to find, for each input value x, a confidence interval containing $y(x)$ with confidence level $1-\alpha$. This is done via probability possibility transformations [22]. Varying α leads to a bunch of nested intervals that can be modelled by fuzzy intervals faithful to the dispersion of the y_i's in the vicinity of each input data x_i.

The last approach we can think of is sensitivity analysis yielding all regression results one would obtain from all precise datasets $\mathbf{d}$ consistent with $\mathbb{D}$. Strangely enough this technique is seldom proposed. The aim is to find

the range of results one would have obtained with linear regression, had the data been precise. Formally it can be posed as follows: Find

$$Y(x) = \{\hat{a}(\mathbf{d})x + \hat{b}(\mathbf{d}) : \mathbf{d} \in \mathbb{D}\}$$
$$= \{\hat{a}x + \hat{b}, \forall \hat{a}, \hat{b} \text{ that minimize } \sum_{i=1}^{n}(ax_i + b - y_i)^2, \forall y_i \in [\underline{y}_i, \overline{y}_i], i = 1, ...n\}$$

It is clear that the envelope of the results is a set-valued function $Y(x)$ that has little chance of being defined by affine upper and lower bounds. This approach, which can genuinely be called epistemic regression has been recently applied to kriging in geostatistics [36, 37].

8 Conclusion

This position paper has argued that the use of set-valued and fuzzy mathematics in information processing tasks gives the opportunity to reason about knowledge, an issue not so popular in data-driven studies. However, one should distinguish between genuine set-valued problems where sets stand for existing entities and epistemic data analysis problems where sets represent incomplete information. This distinction impacts the very way new problems can be posed so as to be meaningful in practice. Adding knowledge representation and reasoning to the modeling paradigm seems to be a good way to reconcile Artificial Intelligence and numerical engineering methods.

Strangely enough fuzzy set-based information processing techniques gathered under the Soft Computing flag are not set-valued methods, as they aim most of the time at computing standard numerical functions using fuzzy rules and neural networks, exploiting stochastic metaheuristics to optimise the fit. A fuzzy system is then seldom viewed as an epistemic fuzzy set of systems. Adopting the latter view could lead to fruitful developments of fuzzy sets methods in a direction not yet much considered in the engineering sciences, beyond rehashing good old fuzzy rule-based systems further.

References

1. Baudrit C, Couso I, Dubois D (2007) Joint propagation of probability and possibility in risk analysis: Towards a formal framework. *Int J Approx Reas* 45:82–105
2. Bertoluzza ASC, Salas A, Corral N (1995) On a new class of distances between fuzzy numbers. *Mathware and Soft Comp* 2:71–84
3. Boukezzoula R, Galichet S, Bisserier A (2011) A Midpoint-Radius approach to regression with interval data. *Int. J. Approx. Reasoning* 52:1257–1271
4. Bouyssou D, Dubois D, Pirlot M, Prade H, eds. (2009) *Decision-making Process — Concepts and Methods*. ISTE London & Wiley, New-York
5. Colubi A (2009) 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

6. Colubi A, González-Rodríguez G, Gil MA, Trutschnig W (2011) Nonparametric criteria for supervised classification of fuzzy data. *Int. J. Approx Reas* 52:1272–1282
7. Couso I, Dubois D (2009) On the Variability of the Concept of Variance for Fuzzy Random Variables. *IEEE Trans Fuzzy Syst* 17:1070–1080
8. Couso I, Sánchez L (2011) Upper and lower probabilities induced by a fuzzy random variable. *Fuzzy Sets Syst* 165:1–23
9. De Campos LM, Lamata MT, Moral S (1990) The concept of conditional fuzzy measure. *Int. J. of Intell Syst* 5:237–246
10. De Cooman G, Walley P (2002) An imprecise hierarchical model for behaviour under uncertainty. *Theory and Decision* 52:327–374
11. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. *Ann Math Stat* 38:325–339
12. Diamond P (1988) Fuzzy least squares. *Inform Sci* 46:141–157
13. Diamond P, Kloeden P (1994) Metric spaces of fuzzy sets. World Scientific, Singapore
14. Dubois D (2006) Possibility theory and statistical reasoning. *Comp Stat & Data Anal* 51:47–69
15. Dubois D (2011) The role of fuzzy sets in decision sciences: Old techniques and new directions. *Fuzzy Sets Syst* 184:3–28
16. Dubois D, Prade H (1988) *Possibility Theory*. Plenum Press, New York
17. Dubois D, Prade H (1988) Incomplete conjunctive information. *Comp & Math Appl* 15:797–810
18. Dubois D, Prade H (1992) When upper probabilities are possibility measures. *Fuzzy Sets Syst* 49:65–74
19. Dubois D, Prade H (1997) The three semantics of fuzzy sets. *Fuzzy Sets Syst* 90:141–150
20. Dubois D, Prade H (2009) Formal representations of uncertainty. Chap. 3 in [4], 85–156
21. Dubois D, Prade H (2012) Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets. *Fuzzy Sets Syst* 192:3–24
22. Dubois D, Foulloy L, Mauris G, Prade H (2004) Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing* 10:273–297
23. Fagin R, Halpern JY (1991) A new approach to updating beliefs. In: Bonissone PP, Henrion M, Kanal LN, Lemmer JF (eds.) *Uncertainty in Articial Intelligence (UAI'91)*, 347–374. Elsevier, New York
24. Ferraro MB, Coppi R, González-Rodríguez G, Colubi A (2010) A linear regression model for imprecise response. *Int J Approx Reas* 51:759–770
25. Ferson S, Ginzburg L, Kreinovich V, Longpre L, Aviles M (2002) Computing variance for interval data is NP-hard. *ACM SIGACT News* 33:108–118
26. González-Rodríguez G, Blanco A, Colubi A, Lubiano MA (2009) Estimation of a simple linear regression model for fuzzy random variables. *Fuzzy Sets and Systems* 160(3):357–370
27. González-Rodríguez G, Colubi A, Gil MA (2012) Fuzzy data treated as functional data. A one-way ANOVA test approach. *Comp Stat and Data Anal* 56(4):943–955
28. Körner R (1997) On the variance of fuzzy random variables. *Fuzzy Sets Syst* 92:83–93
29. Halpern JY, Fagin R, Moses Y, Vardi MY (2003) *Reasoning About Knowledge*. MIT Press, Cambridge
30. Herzig A, Lang J, Marquis P (2003) Action representation and partially observable planning using epistemic logic. *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI-03)*, 1067–1072. Morgan Kaufmann, San Francisco
31. Kendall DG (1974) Foundations of a theory of random sets. In: Harding EF, Kendall DG, eds. *Stochastic geometry*, 322–376 J. Wiley & Sons, New York
32. Kruse R, Meyer K (1987) *Statistics with Vague Data*. D. Reidel, Dordrecht
33. Kwakernaak H (1978) Fuzzy random variables — I. definitions and theorems. *Inform Sci* 15:1–29

34. Kwakernaak H (1979) Fuzzy random variables — II. Algorithms and examples for the discrete case. *Inform Sci* 17:253–278
35. Lindley DV (1982) Scoring rules and the inevitability of probability. *Int Statist Rev* 50:1–26
36. Loquin K, Dubois D (2010) Kriging and epistemic uncertainty: a critical discussion. In: Jeansoulin R *et al.* (eds). *Methods for Handling Imperfect Spatial Information*, 269–305. Springer, Berlin Heidelberg New York
37. Loquin K, Dubois D (2010) Kriging with ill-known variogram and data. *Scalable Uncertainty Management (SUM 2010)*, LNAI 6379:219–235. Springer, Berlin Heidelberg New York
38. Matheron G (1975) *Random Sets and Integral Geometry*. J. Wiley & Sons, New York, NY, USA
39. Moore R (1979) Methods and Applications of Interval Analysis. *SIAM Studies in Applied Mathematics*. SIAM, Philadelphia
40. Nguyen HT (1978) On random sets and belief functions. *J Math Anal Appl* 65:531–542
41. Pichon F, Dubois D, Denoeux T (2012) Relevance and truthfulness in information correction and fusion. *Int J Approx Reas* 53:159–175
42. Prade H, Serrurier M (2011) Maximum-likelihood principle for possibility distributions viewed as families of probabilities. *Proc. IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2011)*, 2987–2993. IEEE Press, Piscataway
43. Puri M, Ralescu D (1986) Fuzzy random variables. *J Math Anal Appl* 114:409–422
44. Ramos-Guajardo AB, Lubiano MA (2012) K-sample tests for equality of variances of random fuzzy sets. *Comp Stat & Data Anal* 56:956–966
45. Shackle GLS (1961) *Decision, Order and Time in Human Affairs (2nd edition)*. Cambridge University Press
46. Shafer G (1976) *A Mathematical Theory of Evidence*. Princeton University Press
47. Shafer G, Tversky A (1985) Languages and designs for probability. *Cogn Sci* 9:309–339
48. Spadoni M, Stefanini L (2011) Computing the variance of interval and fuzzy data. *Fuzzy Sets Syst* 165:24–36
49. Smets P (1997) The normative representation of quantified beliefs by belief functions. *Artif Intell* 92:229–242
50. Tanaka H, Guo P (1999) *Possibilistic Data Analysis for Operations Research*. Physica-Verlag, Heidelberg
51. Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Inform Sci* 179:3964–3972
52. Walley P (1991) *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall
53. Yager RR (1987) Set-based representations of conjunctive and disjunctive knowledge. *Inform Sci* 41:1–22
54. Yen J (1990) Generalizing the Dempster-Shafer theory to fuzzy sets. *IEEE Trans Syst Man and Cybern* 20:559–569
55. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning, part I. *Inform Sci* 8:199–249
56. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst* 1:1–28
57. Zadeh LA (1978) PRUF — a meaning representation language for natural languages. *Int J Man-Mach Stud* 10:395–460

Pricing of Catastrophe Bond in Fuzzy Framework

Piotr Nowak[1] and Maciej Romaniuk[1]

Abstract In the paper we consider catastrophe bonds with a stepwise payoff structure. We use the martingale method to price it under the condition of no arbitrage. We assume a stochastic form of the spot interest rate, replicability of interest rate changes by financial instruments existing in the market as well as independence between a catastrophe occurrence and behaviour of the financial market. The fuzzy sets approach, presented in the paper, may incorporate expertise knowledge to overcome lack of precise data in the discussed case.

1 Introduction

During last years, the insurance industry faced overwhelming risks caused by natural catastrophes. Losses from a single catastrophic event could reach tens of billions $ (see e.g. [14]). The classical insurance mechanisms are not adequate to deal with extreme losses caused by natural catastrophes, because the sources of such losses are strongly dependent in terms of time and localization. Additionally, losses from such events are extremely huge. In contrary, classical insurance approach assumes that losses are modelled by independent and identically distributed (iid) random variables. Additionally, it is commonly assumed that the value of each claim is small comparing to the whole portfolio of insurer. Therefore it may be useful to use new kinds of insurance instruments.

Because daily fluctuations on worldwide financial markets reach also the same scale as losses from natural catastrophes, securitization of losses (i.e. "packaging" losses into form of tradable assets) may be helpful for dealing with results of extreme natural catastrophes (see e.g. [4, 10]). One possible

[1] Systems Research Institute PAS, ul. Newelska 6, 01–447 Warszawa, Poland,
pnowak@ibspan.waw.pl · mroman@ibspan.waw.pl

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 137–150.
springerlink.com

instruments of this type is known as catastrophe bond (Act-of-God bond, cat bond, see e.g. [7]).

Pricing of catastrophe bonds is not a widely discussed subject in mathematical finance. However, many authors emphasize advantages of cat bonds. In financial literature one can find simplified pricing methods (see [1, 8]) or other approaches, which also have their limitations (see e.g. [8]) or are not strictly mathematically oriented (see e.g. [23]). There are several advanced stochastic pricing models in discrete and continuous time. In some of them a utility function is incorporated to the pricing model (see [3, 5, 21]). However, choosing a well-suited utility function can be an additional problem in practice. An interesting approach was presented by Vaugirard in [22]. The author applied the martingale method for cat bonds pricing. He overcame the problem of non completeness of the market and non-traded insurance-linked underlyings in Merton's manner (see [13]).

This paper is dedicated to the problem of catastrophe bond pricing with a stepwise payoff function. We continue and extend the Vaugirard's approach. Analogously to the author, we use the martingale method. We find pricing formulas, considering the Vasicek and the Merton risk-free spot interest rate dynamics. We assume no arbitrage, replicability of interest rate changes by financial instruments existing in the market as well as independence between catastrophe occurrence and behaviour of the financial market. There is a need to take into account possible errors and uncertainties which arise from estimation of the financial market parameters. In order to price catastrophe bond in case of lack of precise data, we apply fuzzy parameters of the spot interest rate processes. Then the Monte Carlo simulations based on the obtained fuzzy pricing formulas are carried out. In the case of the Vasicek interest rate model, we continue our considerations from [16]. However, in this paper, real world data describing catastrophic events are used.

This paper is organized as follows. In Section 2 we present some preliminaries. In Section 3 we discuss features of catastrophe bonds, especially with a stepwise payoff function. In Section 4 we present the general cat bond pricing formula and we price the considered type of catastrophe bond. In Section 5 we conduct simulations in order to find appropriate prices for the fuzzy approach. We conclude the paper in Section 6.

2 Preliminaries

2.1 Fuzzy and Interval Arithmetic

In this section we recall some basic facts about fuzzy sets and numbers.

Let $\tilde{A}$ be a fuzzy subset of $\mathbb{R}$. We denote by $\mu_{\tilde{A}}$ its membership function $\mu_{\tilde{A}} : \mathbb{R} \to [0,1]$, and by $\tilde{A}_\alpha = \{x : \mu_{\tilde{A}} \geq \alpha\}$ the α-level set of $\tilde{A}$, where $\tilde{A}_0$ is the closure of the set $\{x : \mu_{\tilde{A}} > 0\}$.

Let $\tilde{a}$ be a fuzzy number (in particular let $\mu_{\tilde{a}}$ be upper semicontinuous). Then the α-level set $\tilde{a}_\alpha$ is a closed interval by definition, which can be denoted by $\tilde{a}_\alpha = [\tilde{a}_\alpha^L, \tilde{a}_\alpha^U]$ (see e.g. [26]).

We recall the arithmetic of fuzzy numbers. Let $\odot$ be a binary operator $\oplus$, $\ominus$, $\otimes$ or $\oslash$ between fuzzy numbers $\tilde{a}$ and $\tilde{b}$, where the binary operators correspond to $\circ$: $+$, $-$, $\times$ or $/$, according to the "Extension Principle" (see [26]). Let $\odot_{int}$ be a binary operator $\oplus_{int}$, $\ominus_{int}$, $\otimes_{int}$ or $\oslash_{int}$ between two closed intervals $[a,b]$ and $[c,d]$.

Then $[a,b] \odot_{int} [c,d] = \{z \in \mathbb{R} : z = x \circ y, x \in [a,b], y \in [c,d]\}$, where $\circ$ is usual operation $+, -, \times$ and $/$, if the interval $[c,d]$ does not contain zero in the latter case. Therefore, if $\tilde{a}, \tilde{b}$ are fuzzy numbers, then $\tilde{a} \odot \tilde{b}$ is also a fuzzy number and defined via its α-level sets by

$$(\tilde{a} \oplus \tilde{b})_\alpha = \tilde{a}_\alpha \oplus_{int} \tilde{b}_\alpha = [\tilde{a}_\alpha^L + \tilde{b}_\alpha^L, \tilde{a}_\alpha^U + \tilde{b}_\alpha^U] \,,$$

$$(\tilde{a} \ominus \tilde{b})_\alpha = \tilde{a}_\alpha \ominus_{int} \tilde{b}_\alpha = [\tilde{a}_\alpha^L - \tilde{b}_\alpha^U, \tilde{a}_\alpha^U - \tilde{b}_\alpha^L] \,,$$

$$(\tilde{a} \otimes \tilde{b})_\alpha = \tilde{a}_\alpha \otimes_{int} \tilde{b}_\alpha =$$

$$= [\min\{\tilde{a}_\alpha^L \tilde{b}_\alpha^L, \tilde{a}_\alpha^L \tilde{b}_\alpha^U, \tilde{a}_\alpha^U \tilde{b}_\alpha^L, \tilde{a}_\alpha^U \tilde{b}_\alpha^U\}, \max\{\tilde{a}_\alpha^L \tilde{b}_\alpha^L, \tilde{a}_\alpha^L \tilde{b}_\alpha^U, \tilde{a}_\alpha^U \tilde{b}_\alpha^L, \tilde{a}_\alpha^U \tilde{b}_\alpha^U\}] \,,$$

$$(\tilde{a} \oslash \tilde{b})_\alpha = \tilde{a}_\alpha \oslash_{int} \tilde{b}_\alpha =$$

$$= [\min\{\tilde{a}_\alpha^L / \tilde{b}_\alpha^L, \tilde{a}_\alpha^L / \tilde{b}_\alpha^U, \tilde{a}_\alpha^U / \tilde{b}_\alpha^L, \tilde{a}_\alpha^U / \tilde{b}_\alpha^U\}, \max\{\tilde{a}_\alpha^L / \tilde{b}_\alpha^L, \tilde{a}_\alpha^L / \tilde{b}_\alpha^U, \tilde{a}_\alpha^U / \tilde{b}_\alpha^L, \tilde{a}_\alpha^U / \tilde{b}_\alpha^U\}] \,,$$

if α-level set $\tilde{b}_\alpha$ does not contain zero for all $\alpha \in [0,1]$ in the case of $\oslash$.

2.2 Stochastic and Financial Preliminaries

We begin with notations and basic definitions concerning catastrophe bonds and their pricing.

We apply stochastic models with continuous time and time horizon of the form $[0, T']$, where $T' > 0$. Date T of maturity of catastrophe bonds is not later than T', i.e. $T \leq T'$. We consider two probability measures, P and Q, and we denote by E^P and E^Q the expectations with respect to them. We define stochastic processes and random variables with respect to P.

Let $(W_t)_{t \in [0,T']}$ be a Brownian motion. It will be used in the stochastic model of the risk-free interest rate. Let $(U_i)_{i=1}^{\infty}$ be a sequence of identically distributed random variables. We treat U_i as total loss caused by the i-th catastrophic event. We also define compound Poisson process by formula

$$\tilde{N}_t = \sum_{i=1}^{N_t} U_i,\ t \in [0, T'],$$

where N_t is a Poisson process with an intensity $\kappa > 0$.

For each $t \in [0, T']$ the integer N_t is equal to the number of catastrophic events till the moment t. Moments of jumps of process $(N_t)_{t\in[0,T']}$ are interpreted as moments of catastrophic events.

For each $t \in [0, T']$ process $\tilde{N}_t$ describes the aggregated catastrophe losses till the moment t. $\left(\tilde{N}_t\right)_{t\in[0,T']}$ is a nondecreasing stochastic process, with right–continuous trajectories of a stepwise form. Heights of its jumps are equal to values of losses during catastrophic events.

All the above processes and random variables are defined on a filtered probability space $\left(\Omega, F, (F_t)_{t\in[0,T']}, P\right)$. The filtration $(F_t)_{t\in[0,T']}$ is given by formula

$$F_t = \sigma\left(F_t^0 \cup F_t^1\right)\ ,\ F_t^0 = \sigma\left(W_s, s \le t\right), F_t^1 = \sigma\left(\tilde{N}_s, s \le t\right), t \in [0, T'].$$

We assume that $F_0 = \sigma(\{A \in F : P(A) = 0\})$ and that $(W_t)_{t\in[0,T']}$, $(N_t)_{t\in[0,T']}$ and $(U_i)_{i=1}^{\infty}$ are independent. Then the probability space with filtration satisfies standard assumptions, i.e. σ-algebra F is P-complete, filtration $(F_t)_{t\in[0,T']}$ is right–continuous and F_0 contains all the sets from F of P-probability zero. Moreover, we assume that the random variables U_i, $i = 1, 2, ...$ have bounded second moment.

We denote by $(B_t)_{t\in[0,T']}$ the banking account satisfying equation $dB_t = r_t B_t dt$, $B_0 = 1$, where $r = (r_t)_{t\in[0,T']}$ is a risk-free spot interest rate.

Let us assume that zero-coupon bonds are traded in the market. We denote by $B(t, T)$ the price of a zero-coupon bond with maturity T at time t and with the face value (principal) equal to 1.

We price catastrophe bonds under the assumption of no possibility of arbitrage in the market. Let us make two additional assumptions. We first assume that investors are neutral toward nature jump risk. This assumption has practical confirmations in the market (see e.g. [1, 22]). Secondly, we assume that changes in interest rate r can be replicated by existing financial instruments (especially zero-coupon bonds).

3 Catastrophe Bonds

As it was mentioned before, there are many problems with classical insurance mechanisms and alternative financial or insurance instruments may be useful by insurers. One of the most popular catastrophe-linked security is the catastrophe bond, known also as *cat bond* or *Act-of-God* bond (see [7, 9, 17]).

There is one important difference among cat bonds and more classical types of bonds. The payment function of a cat bond depends on some additional random variable, i.e. occurrence of some natural catastrophe in the specified region and the fixed time interval. Such an event is called *triggering point*. Catastrophe bonds may be related to various kinds of triggering points — e.g. to magnitudes of earthquakes, the losses from flood, the insurance industry index of losses, some parameters of catastrophe events, etc. The structure of payments for cat bonds depends also on some primary underlying asset, like LIBOR (London Interbank Offered Rate).

3.1 Catastrophe Bond with Stepwise Payoff Function

Let $0 < K_1 < ... < K_n$, $n > 1$, be a sequence of constants and $\tau_i : \Omega \to [0, T']$, $1 \leq i \leq n$ be a sequence of stopping times defined as follows

$$\tau_i(\omega) = \inf_{t \in [0,T']} \left\{ \tilde{N}_t(\omega) > K_i \right\} \wedge T', \ 1 \leq i \leq n.$$

Let $w_1 < w_2 < ... < w_n$ be a sequence of nonnegative constants, for which $\sum_{i=1}^{n} w_i \leq 1$. Let $\Phi = \sum_{i=1}^{n} w_i \Phi_i$, where Φ_i are cumulative distribution functions of τ_i.

Definition 1. We denote by $IB_s(T, FV)$ a catastrophe bond satisfying the following assumptions:

a) If the catastrophe does not occur in the period $[0, T]$, i.e. $\tau_1 > T$, the bondholder is paid the face value FV;
b) If $\tau_n \leq T$, the bondholder receives the face value minus the sum of write-down coefficients in percentage $\sum_{i=1}^{n} w_i$.
c) If $\tau_{k-1} \leq T < \tau_k$, $1 < k \leq n$, the bondholder receives the face value minus the sum of write-down coefficients in percentage $\sum_{i=1}^{k-1} w_i$.
d) Cash payments are done at date of maturity T.

4 Pricing of Catastrophe Bonds

4.1 General Formula

The first step in our considerations is to obtain the valuation formula for $IB_s(T, FV)$. We denote by $\nu_{IB_{cat}(T,FV)}$ a general payoff function which depends on T, FV and the compound Poisson process $\tilde{N}$. The following theorem from [18] for a general form of catastrophe bond $IB_{cat}(T, FV)$ with a payoff function $\nu_{IB_{cat}(T,FV)}$ is applied.

Theorem 1. *Let $IB(t)$ be the price of a $IB_{cat}(T, FV)$ at time t. Then*

$$IB(t) = E^Q\left(\exp\left(-\int_t^T r_u du\right)\nu_{IB_{cat}(T,FV)}|F_t\right). \tag{1}$$

In particular,

$$IB(0) = E^Q\left(\exp\left(-\int_0^T r_u du\right)\right)E^Q\nu_{IB_{cat}(T,FV)}. \tag{2}$$

In the above theorem measure Q is defined by the Radon-Nikodym derivative:

$$\frac{dQ}{dP} = \exp\left(-\int_0^T \lambda_u dW_u - \frac{1}{2}\int_0^T \lambda_u^2 du\right) P\text{-a.s.}$$

for a predictable process λ_u, connected with risk premium for risk-free bonds. For Q the family $B(t,T)$, $t \le T \le T'$, is an arbitrage-free family of zero-coupon bond prices with respect to r, i.e. for each $T \in [0,T']$ $B(T,T) = 1$ and the process of discounted zero-coupon bond price $B(t,T)/B_t$, $t \in [0,T]$, is a martingale with respect to Q.

4.2 The Vasicek Interest Rate Model

We show crisp and fuzzy pricing formulas, introduced and proved by us in [16, 18], for cat bonds with the stepwise form of payoff function and the Vasicek risk-free spot interest rate model, described by the following equation

$$dr(t) = a(b - r(t))dt + \sigma dW_t \tag{3}$$

for positive constants a, b and σ. The Vasicek model is very popular and often used for modeling the risk-free interest rate in the market.

The following theorem (proved in [18]) gives the pricing formula for $IB_s(T, FV)$.

Theorem 2. *Let $IB(0)$ be the price of an $IB_s(T, FV)$ at time 0.*

$$IB(0) = FV \cdot e^{-T\cdot R(T,r(0))}\{1 - \Phi(T)\}, \tag{4}$$

where $R(\theta, r) = R_\infty - \frac{1}{a\theta}\left\{(R_\infty - r)\left(1 - e^{-a\theta}\right) - \frac{\sigma^2}{4a^2}\left(1 - e^{-a\theta}\right)^2\right\}$
and $R_\infty = b - \frac{\lambda\sigma}{a} - \frac{\sigma^2}{2a^2}$.

We introduce fuzzy numbers $\tilde{a}$, $\tilde{b}$, $\tilde{\sigma}$ and $\tilde{r}_0$ in place of a, b, σ and r_0 to model the uncertainty in the market. We treat market price of risk parameter

as a fuzzy number $\tilde{\lambda}$. We have the following theorem (proved in [16] using Theorem 2 and then extended via the extension principle).

Theorem 3.

$$\tilde{IB}(0) = FV \otimes e^{-T\otimes\tilde{R}(T)} \otimes \{1 - \Phi(T)\}, \tag{5}$$

where

$$\tilde{R}(T) = \tilde{R}_\infty \ominus \left\{ \left(\tilde{R}_\infty \ominus \tilde{r}_0 \right) \otimes \left(1 \ominus e^{-\tilde{a}\otimes T} \right) \ominus \tilde{\sigma}\otimes\tilde{\sigma} \otimes \left(1 \ominus e^{-\tilde{a}\otimes T} \right) \right. \\ \left. \otimes \left(1 \ominus e^{-\tilde{a}\otimes T} \right) \oslash (4 \otimes \tilde{a}\otimes\tilde{a}) \right\} \oslash (\tilde{a} \otimes T) \tag{6}$$

and $\tilde{R}_\infty = \tilde{b} \ominus \tilde{\lambda}\otimes\tilde{\sigma} \oslash \tilde{a} \ominus \tilde{\sigma}\otimes\tilde{\sigma} \oslash (2\otimes\tilde{a}\otimes\tilde{a})$.

It is possible to calculate the α-level sets of $\tilde{IB}(0)$. However, since their form is relatively complex in the considered case, we replace their calculation by Monte Carlo simulations, conducted in Section 5.

4.3 The Merton Interest Rate Model

We consider the Merton dynamics of risk-free spot interest rate $(r_t)_{t\in[0,T']}$. The interest rate behaviour is described by the following equation

$$dr_t = \mu dt + \sigma dW_t, t \in [0, T'],$$

for constants μ and $\sigma > 0$. Its solution is of the form

$$r_t = r_0 + \mu t + \sigma W_t, t \in [0, T'],$$

where $r_0 \geq 0$. This model, introduced by Merton in 1973, was one of the first widely used stochastic models of the interest rate. Our aim is to introduce and prove the valuation formula for catastrophe bond $IB_s(T, FV)$, analogous to before, assuming the Merton interest rate dynamics. The following theorem gives the pricing formula in the crisp case.

Theorem 4. *Let $IB_s(0)$ be the price at the moment zero 0 of the cat bond $IB_s(T, FV)$ for the Merton model of the spot interest rate. Then*

$$IB_s(0) = FV \cdot e^{R(T)} \{1 - \Phi(T)\}, \tag{7}$$

where

$$R(T) = -r_0 T - \frac{(\mu - \lambda\sigma) T^2}{2} + \frac{\sigma^2 T^3}{6}. \tag{8}$$

Sketch of the proof. From Theorem 1 it follows that

$$IB\left(0\right)=E^{Q}\left(\exp\left(-\int_{0}^{T}r_{u}du\right)\right)FV\cdot E^{Q}\left\{1-\sum_{i=1}^{n}w_{i}I_{\tau_{i}\leq T}\right\}.$$

From the zero-coupon bond pricing formula for the Merton interest rate model (see e.g. [15]) it follows that

$$E^{Q}\left(\exp\left(-\int_{0}^{T}r_{u}du\right)\right)=e^{R(T)}.$$

Since τ and W are independent, $I_{\tau_i\leq T}$ and $\frac{dQ}{dP}$ are independent. Therefore

$$E^{Q}\left\{1-\sum_{i=1}^{n}w_{i}I_{\tau_{i}\leq T}\right\}=1-\sum_{i=1}^{n}w_{i}E^{P}\left(I_{\tau_{i}\leq T}\right)=1-\Phi\left(T\right).$$

Finally, the pricing formula at time $t=0$ has the form (7). □

Let $\mathcal{F}\left(\mathbb{R}\right)$ be the set of all fuzzy numbers. The proposition below was proved in [24].

Proposition 1. *Let $f:\mathbb{R}\rightarrow\mathbb{R}$ be a function and $\{x:r=f(x)\}$ is a compact set for all r. Then f induces a fuzzy-valued function $\tilde{f}:\mathcal{F}\left(\mathbb{R}\right)\rightarrow\mathcal{F}\left(\mathbb{R}\right)$ via the extension principle and for each $\tilde{\Lambda}\in\mathcal{F}\left(\mathbb{R}\right)$ the α-level set of $\tilde{f}(\tilde{\Lambda})$ is $\tilde{f}(\tilde{\Lambda})_{\alpha}=\{f(x):x\in\tilde{\Lambda}_{\alpha}\}$.*

In this proposition the assumption of continuity is not necessary, however in this paper it will be applied only for the exponential function (i.e. in proof of Theorem 5).

Our aim is to price catastrophe bond in case when the parameters of the spot interest rate are not precisely known. To model this uncertainty we introduce fuzzy numbers $\tilde{\mu}$, $\tilde{\sigma}$ and $\tilde{r}_0$ in place of μ, σ and r_0. We also treat the market price of risk parameter as a fuzzy number. Therefore we replace the parameter λ by its fuzzy counterpart $\tilde{\lambda}$. We assume that $\tilde{\sigma}$ and $\tilde{r_0}$ are non-negative fuzzy numbers, i.e. their membership funcions are equal to 0 for all negative arguments. The following theorem is the fuzzy version of the pricing formula.

Theorem 5. *The price of the catastrophe bond is given by*

$$I\tilde{B}_{s}\left(0\right)=FV\otimes e^{\tilde{R}(T)}\otimes\left\{1-\Phi\left(T\right)\right\},\tag{9}$$

where

$$\tilde{R}\left(T\right)=\tilde{r}_{0}\otimes\left(-T\right)\oplus\left(\tilde{\mu}\ominus\tilde{\lambda}\otimes\tilde{\sigma}\right)\otimes\frac{-T^{2}}{2}\oplus\tilde{\sigma}\otimes\tilde{\sigma}\otimes\frac{T^{3}}{6}.\tag{10}$$

Moreover, for $\alpha\in[0,1]$

$$\left(I\tilde{B}_s\left(0\right)\right)_\alpha = \left[FV\left\{1-\Phi\left(T\right)\right\}e^{\left(\tilde{R}(T)\right)_\alpha^L}, FV\left\{1-\Phi\left(T\right)\right\}e^{\left(\tilde{R}(T)\right)_\alpha^U}\right], \quad (11)$$

where

$$\left(\tilde{R}\left(T\right)\right)_\alpha^L = -\left(\tilde{r}_0\right)_\alpha^U T - \left(\tilde{\mu}_\alpha^U - \left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^L\right)\frac{T^2}{2} + \left(\tilde{\sigma}_\alpha^L\right)^2\frac{T^3}{6} \quad (12)$$

and

$$\left(\tilde{R}\left(T\right)\right)_\alpha^U = -\left(\tilde{r}_0\right)_\alpha^L T - \left(\tilde{\mu}_\alpha^L - \left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^U\right)\frac{T^2}{2} + \left(\tilde{\sigma}_\alpha^U\right)^2\frac{T^3}{6}. \quad (13)$$

Proof. Replacing crisp parameters by their fuzzy counterparts and arithmetic operators $+, -, .$ by $\oplus, \ominus, \otimes$ in (7) and (8), we obtain formulas (9) and (10). Let $\alpha \in [0,1]$. For a given fuzzy number $\tilde{A}$ its α-level set is denoted similarly as in Section 2.1. Then $\left(\tilde{r}_0 \otimes (-T)\right)_\alpha = \left[-\left(\tilde{r}_0\right)_\alpha^U T, -\left(\tilde{r}_0\right)_\alpha^L T\right]$. Since $\tilde{\sigma}$ is non-negative, $\left(\tilde{\lambda}\otimes\tilde{\sigma}\right)_\alpha = \left[\left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^L, \left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^U\right]$, $\left(\tilde{\mu}\ominus\tilde{\lambda}\otimes\tilde{\sigma}\right)_\alpha = \left[\tilde{\mu}_\alpha^L - \left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^U, \tilde{\mu}_\alpha^U - \left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^L\right]$, $\left(\left(\tilde{\mu}\ominus\tilde{\lambda}\otimes\tilde{\sigma}\right)\otimes\frac{-T^2}{2}\right)_\alpha = \left[-\left(\tilde{\mu}_\alpha^U - \left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^L\right)\frac{T^2}{2}, -\left(\tilde{\mu}_\alpha^L - \left(\tilde{\lambda}_\alpha \otimes_{int} \tilde{\sigma}_\alpha\right)^U\right)\frac{T^2}{2}\right]$.

Furthermore, $\left(\tilde{\sigma}\otimes\tilde{\sigma}\otimes\frac{T^3}{6}\right)_\alpha = \left[\left(\tilde{\sigma}_\alpha^L\right)^2\frac{T^3}{6}, \left(\tilde{\sigma}_\alpha^U\right)^2\frac{T^3}{6}\right]$. From the above equalities it follows that (12) and (13) hold. Since function exp satisfies the assumptions of Proposition 1 and is increasing, $\left(e^{\tilde{R}(T)}\right)_\alpha = \left[e^{\left(\tilde{R}(T)\right)_\alpha^L}, e^{\left(\tilde{R}(T)\right)_\alpha^U}\right]$ and finally, we obtain (11). □

The α-level set (e.g. $\alpha = 0.95$) of fuzzy price $I\tilde{B}_s\left(0\right)$ can be treated by a financial analyst as an interval of the cat bond prices. Then the financial analyst can pick any value from this interval as the catastrophe bond price with an acceptable membership degree. For example, if the real market price is outside of such the interval, an appropriate course of action (i.e. selling or buying of the asset) may be taken by the decision-maker. Therefore the α-level set can be a comfortable tool for his (her) latter use.

5 Numerical Examples of Fuzzy Approach

In order to find the price of the model of the catastrophe bond described in Section 3.1 according to Theorem 3, the appropriate Monte Carlo simulations are conducted.

First set of parameters is used for modelling of losses. In this case we assume that the quantity of losses is modelled by homogeneous Poisson process

(HPP) with the intensity κ and the value of each loss is given by a random variable from lognormal distribution with parameters μ_{LN} and σ_{LN} which are commonly used in simulation of values of risk events in insurance (see [2]). Also other types of complex probabilistic distributions or simulations based on historical records are possible. We assume that parameters of the value of each loss are the same as described in [2] for natural catastrophic events in the United States provided by ISO's (Insurance Service Office Inc.) Property Claim Services (PCS), therefore $\mu_{LN} = 17.3570, \sigma_{LN} = 1.7643$. Also intensity of HPP is given by parameter fitted in [2], therefore $\kappa = 31.7143$.

Let us denote by $Q_{\text{HPP-LN}}(x)$ the x-th quantile of cumulated value of losses for HPP process (the number of losses) and lognormal distribution (the value of each loss) with the parameters described above.

The second set of parameters is used for modelling of the risk-free spot interest rate trajectories for the Vasicek model. In this case we use fuzzy parameters, i.e. we assume that parameters are described by α-sets which may be derived e.g. from triangular fuzzy numbers or L-R numbers (i.e. Left-Right numbers). We apply parameters specified in [6] for 1-month interbank rate in case of the UK, then $a = 0.0263, b = 0.0988593, \sigma = 0.01, r_0 = 0.1039$. The actual transformation of real market values to the α-sets may be based on expert knowledge. The expert knowing the data, issue an imprecise opinion in the form of L-R numbers instead of one exact estimate. Other approaches like consensus among experts and aggregation of opinions of decision-makers are also possible (see e.g. [11, 12])

In our paper we assume that α-set for each parameter is the appropriate interval containing value of this parameter. Therefore possible uncertainties of behaviour of some market parameters in future may be incorporated.

For each considered example of catastrophe bond the trading horizon is set on 1 year, face value is equal to 1, and we generate 1000 interest rate trajectories for 10000 simulations of the Poisson process.

In case of *Example I – III* we analyse the estimators of the cat bond price if the limits of α-sets are wider for each experiment (i.e. the appropriate α-level is lower). The triggering points are connected with surpassing the limits given by $K_1 = Q_{\text{HPP-LN}}(0.75), K_2 = Q_{\text{HPP-LN}}(0.85), K_3 = Q_{\text{HPP-LN}}(0.95)$. The values of losses coefficients for bond's holder are equal to $w_1 = 0.1$, $w_2 = 0.2$, $w_3 = 0.3$.

Then based on equation (5) we obtain estimators for the price of catastrophe bond presented in Table 1. The average, maximum, 95% quantile, 99% quantile and third quartile increase. Minimum, first quartile, 1% quantile and 5% quantile decrease. In case of fuzzy numbers, minimum and maximum may be seen as the most important estimators of the boundaries of the α-level sets.

Then in *Example IV – VI* we analyse the estimators of catastrophe bond price if the values of μ_{LN} and σ_{LN} are increased, i.e. the expected value of single catastrophe and its variance are higher. We assume that for the given α-level (e.g. 0.95 in our expert's opinion), the α-set are described by intervals

Table 1 Numerical estimators for price of catastrophe bond in *Example I, II, III*

	Example I	Example II	Example III
$\tilde{a}_\alpha$	[0.02,0.03]	[0.018,0.032]	[0.016,0.34]
$\tilde{b}_\alpha$	[0.09,0.1]	[0.088,0.12]	[0.086,0.14]
$\tilde{\sigma}_\alpha$	[0.005,0.015]	[0.003,0.017]	[0.001,0.019]
$(\tilde{r}_0)_\alpha$	[0.1,0.11]	[0.08,0.12]	[0.06,0.14]
Average	0.837184	0.841602	0.841436
First quartile	0.834994	0.833033	0.825564
Median	0.837136	0.841477	0.841171
Third quartile	0.839285	0.849581	0.857395
Standard deviation	0.0027443	0.00968603	0.018983
Minimum	0.829741	0.821644	0.807056
1% quantile	0.831457	0.824369	0.80876
5% quantile	0.832883	0.826635	0.812306
95% quantile	0.841669	0.857042	0.871743
99% quantile	0.842963	0.859555	0.875402
Maximum	0.844653	0.861277	0.876863

$$\tilde{a}_\alpha = [0.022, 0.03]\ ,\ \tilde{b}_\alpha = [0.094, 0.102]\ ,$$
$$\tilde{\sigma}_\alpha = [0.008, 0.012]\ ,\ (\tilde{r}_0)_\alpha = [0.08, 0.12] \quad (14)$$

and the catastrophe bond is the same as in *Example I*. The obtained estimators may be found in Table 2. As we can see, estimators for higher values of μ_{LN} and σ_{LN} are lower and the differences in obtained estimators are significant.

Table 2 Numerical estimators for the price of the catastrophe bond in *Example IV, V, VI*

	Example IV	Example V	Example VI
μ_{LN}	17.357	17.5	17.7
σ_{LN}	1.7643	1.9	2.1
Average	0.841024	0.760878	0.611699
First quartile	0.832836	0.753238	0.606067
Median	0.84049	0.760881	0.611439
Third quartile	0.849353	0.767846	0.617257
Standard deviation	0.00974711	0.0088759	0.00711116
Minimum	0.822918	0.741471	0.594264
1% quantile	0.82414	0.744771	0.598208
5% quantile	0.826309	0.747187	0.600847
95% quantile	0.856755	0.774917	0.62333
99% quantile	0.858714	0.777923	0.625703
Maximum	0.860313	0.780128	0.628003

In case of *Example VII – VIII* we analyse the estimators of the cat bond price for decreasing values of triggering points K_1, K_2, K_3 measured

in $Q_{\text{HPP-LN}}(x)$. We compare results with the *Example IV* mentioned above. We assume that the α-set are described by intervals (14) and the parameters of the catastrophe bond are the same as in *Experiment I*. The obtained estimators may be found in Table 3. As we can see, all of the estimators, including average, are lower for lower values of the triggering points.

Table 3 Numerical estimators for price of catastrophe bond in *Example IV, VII, VIII*

	Example IV	Example VII	Example VIII
K_1	0.75	0.7	0.65
K_2	0.85	0.8	0.75
K_3	0.95	0.9	0.85
Average	0.841024	0.814571	0.787311
First quartile	0.832836	0.806363	0.779829
Median	0.84049	0.814705	0.78694
Third quartile	0.849353	0.822169	0.79477
Standard deviation	0.00974711	0.00944488	0.0090465
Minimum	0.822918	0.794895	0.768637
1% quantile	0.82414	0.797305	0.770891
5% quantile	0.826309	0.800064	0.773088
95% quantile	0.856755	0.82971	0.801741
99% quantile	0.858714	0.832138	0.80403
Maximum	0.860313	0.834588	0.806175

In *Example IX* we find the exact price of the cat bond for the Merton interest rate model according to the formula (11) (see Section 4.3). Monte Carlo method is used only for simulations of catastrophic events losses ("first set of parameters" mentioned earlier) and to calculate the value of function $\Phi(T)$. We apply parameters similar to values from [20] for UK bonds, therefore $(\tilde{r}_0)_\alpha = [0.105, 0.115]$, $\tilde{\mu}_\alpha = [-0.0003, -0.0001]$, $\tilde{\sigma}_\alpha = [0.005, 0.015]$. Additionally, $K_1 = Q_{\text{HPP-LN}}(0.75)$, $K_2 = Q_{\text{HPP-LN}}(0.95)$, $w_1 = 0.2$, $w_2 = 0.3$. Other parameters are the same as in *Experiment I*. In such case, the price calculated according to formula (11) is equal to $\left(I\tilde{B}_s(0)\right)_\alpha = [0.833472, 0.841961]$.

6 Conclusions

In this paper we price catastrophe bond with a stepwise payoff function. The stochastic approach based on hte martingale method is applied. We consider the Vasicek and the Merton interest rate models. Because of possible errors and uncertainties the fuzzy set approach is applied. Then we analyse the output of some numerical experiments for various sets of parameters.

Acknowledgements The research in this paper has been supported by the COST Action IC0702 STSMs.

References

1. Anderson RR, Bendimerad F, Canabarro E, Finkemeier M (2000) Analyzing insurance-linked securities. *Journal of Risk Finance* 1(2):49–78
2. Chernobai A, Burnecki K, Rachev S, Trueck S, Weron R (2005) *Modeling catastrophe claims with left-truncated severity distributions.* HSC Research Reports, HSC/05/01
3. Cox SH, Pedersen HW (2000) Catastrophe Risk Bonds. *North American Actuarial Journal* 4(4):56–82
4. Cummins JD, Doherty N, Lo A (2002) Can insurers pay for the "big one"? Measuring the capacity of insurance market to respond to catastrophic losses. *Journal of Banking and Finance* 26(2-3):557–583
5. Egamia M, Young VR (2008) Indifference prices of structured catastrophe (CAT) bonds. *Insurance: Mathematics and Economics* 42:771–778
6. Episcopos A (2000) Further evidence on alternative continuous time models of the short-term interest rate. *Journal of International Financial Markets, Institutions and Money* 10:199–212
7. Ermolieva T, Romaniuk M, Fischer G, Makowski M (2007) Integrated model-based decision support for management of weather-related agricultural losses. In: Hryniewicz O, Studziński J, Romaniuk M (eds) *Environmental informatics and systems research. Vol. 1: Plenary and session papers – EnviroInfo 2007.* Shaker Verlag
8. Froot KA (2001) The market for catastrophe risk: A clinical examination. *Journal of Financial Economics* 60(2):529–571
9. George JB (1999) Alternative reinsurance: Using catastrophe bonds and insurance derivatives as a mechanism for increasing capacity in the insurance markets. *CPCU Journal* 52(1):50–54
10. Harrington S E, Niehaus G (2003) Capital, corporate income taxes, and catastrophe insurance. *Journal of Financial Intermediation* 12(4):365–389
11. Heilpern S (1997) Representation and application of fuzzy numbers. *Fuzzy Sets and Systems* 91:259–268
12. Kumar A, Karmakar G (2001) Aggregation of opinions using fuzzy numbers. *Int. Journal of Systems Science* 32(12):1399–1411
13. Merton RC (1976) Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3:125–144
14. Muermann A (2008) Market Price of Insurance Risk Implied by Catastrophe Derivatives. *North American Actuarial Journal* 12(3):221–227
15. Munk C (2011) *Fixed Income Modelling.* Oxford University Press
16. Nowak P, Romaniuk M (to be published) Fuzzy Pricing of Catastrophe Bond with a Stepwise Payoff Function. *Conf. papers 10th Int. Workshop on Intuitionistic Fuzzy Sets and Generalized Nets*
17. Nowak P, Romaniuk M, Ermolieva T (2012) Evaluation of Portfolio of Financial and Insurance Instruments: Simulation of Uncertainty. In: Ermoliev Y, Makowski M, Marti K (eds) *Managing Safety of Heterogeneous Systems.* Springer, Berlin Heidelberg New York
18. Nowak P, Romaniuk M (2011) *Pricing and simulations of catastrophe bonds.* Research Report, RB/44/2011, SRI PAS, Warszawa
19. Nowak P, Romaniuk M (2010) Computing option price for Levy process with fuzzy parameters. *European Journal of Operational Research* 201(1):206–210
20. Nowman KB (1997) Gaussian Estimation of Single-Factor Continuous Time Models of the Term Structure of Interest Rates. *The Journal of Finance* 52(4):1695–1706

21. Reshetar G (2008) Pricing of Multiple-Event Coupon Paying CAT Bond. Working Paper. Swiss Banking Institute, University of Zurich
22. Vaugirard VE (2003) Pricing catastrophe bonds by an arbitrage approach. *The Quarterly Review of Economics and Finance* 43:119–132
23. Wang SS (2004) *Geneva Papers: Etudes et Dossiers, special issue on Insurance and the State of the Art in Cat Bond Pricing* 278:19–29
24. Wu H-Ch (2004) Pricing European options based on the fuzzy pattern of Black-Scholes formula. *Computers & Operations Research* 31:1069–1081
25. Wu X (2000) A New Stochastic Duration Based on the Vasicek and CIR Term Structures Theories. *Journal of Business Finance and Accounting* 27:911–932
26. Zadeh LA (1965) Fuzzy sets. *Information and Control* 8:338–353

Convergence of Heuristic-based Estimators of the GARCH Model

Alexandru Mandes[1], Cristian Gatu[1], and Peter Winker[2]

Abstract The GARCH econometric model is able to describe the volatility of financial data under realistic assumptions and the convergence of its theoretical estimators has been proven. However, when data is "unfriendly" maximum likelihood estimators need to be computed by stochastic optimization algorithms in order to avoid local optima attraction basins, and thus, a new source of uncertainty is introduced. A formal framework for joint convergence analysis of both, the estimators and the heuristic, has been previously described within the context of the GARCH(1,1) model. The aim of this contribution is to adapt and extend this research to asymmetric and multiple lagged GARCH models. Aspects of subset model selection are also investigated.

1 Introduction

Volatility is a way of measuring risk, and therefore it represents a fundamental concept of the financial theory with various applications (e.g., value-at-risk simulations, portfolio risk and management, option pricing models). Volatility is an unobservable phenomenon which can be measured and forecast only within the context of a defined statistical model. Specifically, standard volatility estimation is based on variance or standard deviation. These statistics are sufficient risk measures when the observations are normal, identically and independently distributed. Unfortunately, these assumptions are not realistic for most of the financial data, which usually exhibit features such as: lep-

[1] Alexandru Ioan Cuza University, General Berthelot 16, Iasi, Romania,
alex.mandes@gmail.com, cgatu@info.uaic.ro
[2] Justus Liebig University Giessen, Licher Strasse 64, D-35394 Giessen, and Centre for European Economic Research, Mannheim, Germany,
Peter.Winker@wirtschaft.uni-giessen.de

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 151–163.
springerlink.com

tokurtosis, heteroskedasticity, volatility clustering and leverage effects. Under these conditions, classical volatility indicators have been substituted by more complex econometric models, such as *Generalized Autoregressive Conditional Heteroskedasticity* (GARCH) or *Stochastic Volatility Models*.

The GARCH model was introduced by Engle (1982) and generalized by Bollerslev (1986). The model consists of two linked equations. The first captures the dynamics of returns under the form of an AR process, assuming the possibility of autocorrelation between returns. The second expresses the current variance based on relevant past information: long-term medium value, unexpected return (ARCH factor) and previous value of the variance (GARCH factor). The symmetric normal GARCH(1,1) is the plain vanilla version of the GARCH model and is given by:

$$r_t = c + \epsilon_t \quad \text{where} \quad \epsilon_t | I_{t-1} \sim N(0, \sigma_t^2), \tag{1}$$

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{2}$$

The GARCH parameters are estimated by means of *Maximum Likelihood Estimation* (MLE). The MLE process is based on the construction of a *Log-Likelihood Function* (LLF) which depends on the parametric model that is assumed for the distribution, and on the sample data. It has the following analytical expression:

$$\log L(\psi) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \left(\log(\sigma_t^2) + \left(\frac{\epsilon_t}{\sigma_t} \right)^2 \right), \tag{3}$$

where ψ represents the set of model parameters. In the case of the symmetric GARCH(1,1), $\psi = (c, \omega, \alpha, \beta)$. The LLF is usually maximized through numerical procedures, as there is no analytical solution for the parameters in equation (2) in a heteroskedastic context. In practice, when the assumed model is complex and the data is "unfriendly", these iterative search methods in the solution space are prone to encounter difficulties in identifying the global optimum.

The optimization problem of maximizing (3) can be naturally tackled with heuristic methods, which are designed to avoid the attraction basins of the local optima that occur in a multi-modal hyperspace. *Threshold Accepting* (TA), introduced by Dueck and Scheuer (1990), implements a stochastic search in the neighborhood solution space, represented by vectors of decision variables. The current solution is gradually improved with a decreasing accepted tolerance for worse solutions, which regulates the transition from the exploring to the exploitation phase. Details and the TA pseudocode can be found in [13, 15]. A potential drawback of these heuristic methods is that their stochastic component provides a new source of uncertainty, besides the deviation of the theoretical estimators due to the finiteness of the observation samples. In order to analyze this additional uncertainty, a formal framework for joint convergence analysis of the estimators and the heuristics has been

introduced [15]. TA has been applied to the ML estimates of a GARCH(1,1) model, and the joint convergence property of the optimization algorithm and of the theoretical estimator — when setting the number of iterations of the heuristic as a function of the sample size — has been proven.

The aim of this contribution is to adapt and investigate this previous research to various GARCH extensions. The first, GJR-GARCH, is able to capture market asymmetries by the introduction of an extra parameter which depends on the sign of the previous return. The second one, GARCH(p,q), enables multiple lags for the ARCH and GARCH factors and also allows for holes in the model structure. In the latter case, the optimization problem consists not only in the estimation of the parameters, but also in model selection [11].

2 Convergence of Heuristic-based Estimators

There are two aspects that need to be addressed when analyzing the convergence of the heuristic-based estimators towards the true values of the data generating process (DGP). First, the convergence in probability of the ML theoretical estimators, represented by the vector $\Psi^{\mathrm{ML},T} \in \mathbb{R}^n$, towards their true values $\Psi^{\mathrm{TR}} \in \mathbb{R}^n$, in relation with the sample size T, should be established. For a consistent estimator it holds: for any $\delta > 0,\ \epsilon > 0$ fixed, there is a sample size $T(\delta, \epsilon)$ such that for any $T > T(\delta, \epsilon)$

$$P\left(|\Psi^{\mathrm{ML},T} - \Psi^{\mathrm{TR}}| < \epsilon\right) > 1 - \delta.$$

Second, the heuristic can be interpreted as a stochastic mapping of the search space on a random variable. Thus, based on the distribution of approximations $\Psi^{T,I,r}$ recorded after a number of R restarts ($1 \leq r \leq R$) with I iterations, the asymptotic convergence of the heuristic results towards the theoretical estimator has to be proven. This means that the heuristic can be tuned to approximate the global optimum with an arbitrary accuracy $\epsilon > 0$ and a fixed probability $1 - \delta$ by increasing the number of iterations $I(\delta, \epsilon)$. Put in other words, for any $\epsilon > 0$ and $\delta > 0$, there exists a number of iterations $I_{min} = I(\delta, \epsilon)$ such that for each $I \geq I_{min}$ we find

$$P\left(|\Psi^{T,I,r} - \Psi^{\mathrm{ML},T}| < \epsilon\right) > 1 - \delta.$$

Finally, the two partial results can be combined to obtain the expression of joint convergence of the heuristic-based estimators towards the true values when T goes to infinity and I goes to infinity as a function of T. That is,

$$P\left(|\Psi^{T,I,R} - \Psi^{\mathrm{TR}}| < \epsilon\right) > 1 - \delta,$$

where $\Psi^{T,I,R}$ represents the best result out of the R restarts.

In order to test these relations, Winker and Maringer have considered the GARCH(1,1) model [15]. A number of 100 artificial time series, seeded with the Bollerslev and Ghysels (1996) solution for the DEM/GBP daily exchange rate, from 3rd of January 1984 to 31st of December 1991 have been generated. There were considered six different values for both T and I and, for each of the settings, the heuristic was run for approximately $R \approx 1700$ times. For each time series d and each component of the solution vector p, the mean square deviations between the recorded results and the true parameters (MSD^{TR}), as well as the maximum likelihood parameters (MSD^{ML}), have been computed. These are, respectively,

$$MSD_p^{\mathrm{TR},d,T,I} = \frac{1}{R}\sum_{r=1}^{R}\left(\Psi_p^{d,T,I,r} - \Psi_p^{\mathrm{TR}}\right)^2, \text{ and} \tag{4}$$

$$MSD_p^{\mathrm{ML},d,T,I} = \frac{1}{R}\sum_{r=1}^{R}\left(\Psi_p^{d,T,I,r} - \Psi_p^{\mathrm{ML},d,t}\right)^2. \tag{5}$$

The convergence properties have been evaluated by estimating the linear relationship between the natural logarithms of these dispersion measures and the natural logarithms of the two explaining variables, sample size T and number of iterations I. In order to isolate the estimator and the heuristic effects, the results have been previously grouped by I and respectively by T:

$$\ln\left(MSD_p^{\mathrm{TR},d,T,I}\right) = a_p^{d,I} + b_p^{d,I}\ln(T), \text{ and} \tag{6}$$

$$\ln\left(MSD_p^{\mathrm{ML},d,T,I}\right) = a_p^{d,T} + b_p^{d,T}\ln(I). \tag{7}$$

3 GARCH Extensions

Although simple and restrictive, the symmetric GARCH(1,1) model is widely used in the econometric literature. Brooks (2008) admits that GARCH(1,1) is strong enough to model the volatility bursts in financial data, without the need of additional parameters. Thus, from a parsimonious point of view, it is not always wise to increase the complexity of the model. Even so, theoreticians have suggested over time many variations of the GARCH model with the purpose of including additional features. Two such extensions considered inhere are asymmetric GARCH and generalized GARCH(p,q).

3.1 Asymmetric GARCH

One of the limitations of the GARCH model described in the equations (1)–(2) is the enforcement of a symmetric response to negative and positive market shocks, with the magnitude of the shock as the only input. In practice, it has been noticed that for equities and equity indexes a negative return has a bigger impact on volatility than a positive return of the same size. This asymmetry was explained by the *leverage effect* based on the debt-equity ratio. The opposite asymmetry occurs in the commodity markets. Asymmetries are also common for most types of high frequency financial data which are sensitive to the current state of the market. Various GARCH extensions that include market asymmetries have been previously proposed: A-GARCH (1990) (Engle), exponential GARCH (1991) (Nelson), GJR-GARCH (1993) (Glosten, Jagannathan, Runkle) and Threshold GARCH (1993) (Rabemanajara, Zakoian). A review of these extensions can be found in [1].

Here, GJR-GARCH has been considered as being one of the most complex alternatives. It captures the asymmetry through an extra parameter λ that depends on an indicator function: $1_{\{\epsilon_t<0\}} = 1$ if $\epsilon_t < 0$, and 0 otherwise. The conditional volatility equation (2) is rewritten as follows:

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \lambda 1_{\{\epsilon_{t-1}<0\}}\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2.$$

Parameter estimation is based on the usual normal GARCH likelihood function (3), but σ_t now depends on the additional parameter λ. The TA solution is competitive with the results obtained with commercial software packages such as Matlab and EViews[1]. A comparison of the results obtained on the Bollerslev and Ghysels data is reported in Table 1.

The convergence of the estimator is empirically established using the framework described in section 2. The simulation is based on 50 artificial time series, with the TA solution for the Bollerslev and Ghysels time series in (8) as DGP seed, six different values for sample size T, four different values for the number of iterations I, and 10 restarts for each of the 1,200 combination settings. The true parameter vector is given by:

$$\Psi^{\mathrm{TR}} = [\Psi_0^{\mathrm{TR}} \ldots \Psi_4^{\mathrm{TR}}] = [-0.007864, 0.011786, 0.029426, 0.144650, 0.794912]. \quad (8)$$

Figure 1 illustrates the convergence of the maximum likelihood estimators towards the true values when the sample size increases. For evaluation purposes, the regression parameters in equation (6) have been estimated.

The results are summarized in Table 2(a), where the individual impact of the sample size T is identified by fixing the number of iterations I at different values, as described on the first column. The results imply there is an inverse

[1] The ARMAX/GARCH model parameters have been estimated using the *garchfit* function from Matlab R2008a Econometrics Toolbox, and the EViews 5 built-in GARCH/TARCH equation estimation feature with both, Marquardt and Berndt-Hall-Hall-Hausman, optimization algorithms. The proposed TA has been implemented in Java.

Table 1: GJR-GARCH(1,1) estimators with different implementations and different number of iterations for TA.

Estimators	Implementations					
	EViews		**Matlab**	**Threshold Accepting**		
	Marquardt	BHHH		I=1,000	I=25,000	I=100,000
ψ_0	-	-	-0.0079	-0.0072	-0.0078	-0.0079
ψ_1	0.0099	0.0099	0.0112	0.0178	0.0121	0.0118
$\psi_2(\lambda)$	0.0204	0.0203	0.0283	0.0439	0.0304	0.0294
ψ_3	0.1338	0.1339	0.1405	0.1782	0.1462	0.1446
ψ_4	0.8172	0.8170	0.8014	0.7302	0.7914	0.7949
Conv. rate	0.9510	0.9509	0.9419	0.9083	0.9376	0.9395
LLF	-1106.62[a]	-1106.62[a]	-1106.08[a]	-1107.79	-1106.12	-1106.10

[a] The LLF values of the EViews and Matlab solutions are computed with the TA implementation.

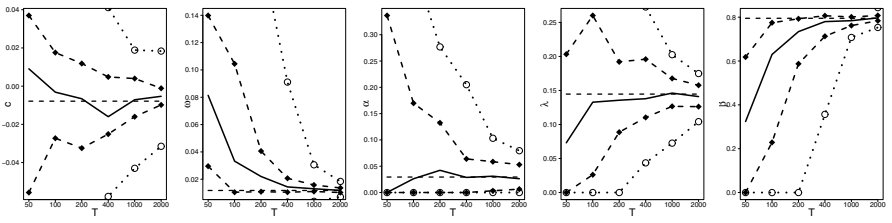


Fig. 1: The convergence of the estimators of 50 time series in relation with T: true values (horizontal dashed line), median (solid line), 25 and 75% quantiles (dashed lines with filled diamonds), minimum and maximum (dotted lines with empty circles).

relation between the dispersion of the estimators measured by MSD^{TR} and the available number of observations T. The asymmetry parameter λ shows some inconsistency, the order of the signed returns playing an important role, besides the sample size. The heuristic optimization convergence can be established by testing the linear relation defined in (7), between $\ln(MSD^{\mathrm{ML}})$ computed as in (5) and $\ln(I)$. The estimators and relevant statistics presented in Table 2(b) show that I has a negative influence on MSD^{ML}. The high values of R^2 prove I to be significant in explaining the convergence of heuristic estimators, including λ, towards the maximum likelihood values.

Figure 2 shows that when the number of iterations is small, the heuristic does not converge for any available sample size. On the other hand, when the number of observations is too small, the heuristic encounters convergence difficulties even for a high number of iterations. The joint convergence of heuristic based estimators towards the true values is assured when the number of iterations I is chosen proportional to the sample size T.

Table 2: Statistics of the regression results for the GJR-GARCH λ estimator convergence.

(a) The influence of T on $MSD^{\mathrm{TR}}(\lambda)$ for fixed I and over d time series.

I	b_p averaged over d	standard deviation of b_p	R^2 averaged over d
1000	-0.267	0.482	0.28
5000	-1.777	10.87	0.45
10000	4.519	110.9	0.44
25000	10.71	170.5	0.43

(b) The influence of I on $MSD^{\mathrm{ML}}(\lambda)$ for fixed T and over d time series.

T	b_p averaged over d	standard deviation of b_p	R^2 averaged over d
50	-0.547	1.079	0.65
100	-0.374	0.340	0.78
200	-0.314	0.610	0.83
400	-0.271	0.118	0.86
1000	-0.260	0.117	0.87
2000	-0.247	0.067	0.89

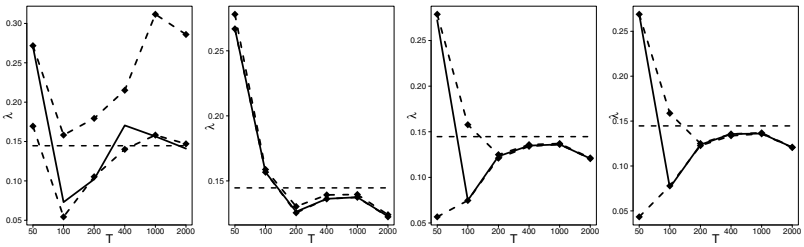


Fig. 2: The convergence of the λ estimator in relation with T for one time series: true values (horizontal dashed line), maximum likelihood (solid line), 10 and 90% quantiles (dashed lines with filled diamonds), for $I = \{1000, 5000, 10000, 25000\}$ (from left to right).

3.2 GARCH(p,q)

GARCH(p,q) allows for additional lags both for the ARCH(p) and for the GARCH(q) factor. The ARCH factor measures the reaction to market shocks, while the GARCH factor measures the persistence of volatility. Engle (2001) recommends this extended version for very long time series, tens of years of daily data or years of intra-day data. In this case, the standard conditional volatility equation (2) becomes:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2. \tag{9}$$

The EViews and Matlab implementations require for the number of lags specified as input and then estimate the parameters of the user-selected model. The objective function is the standard LLF in (3) and the algorithm exhibits an overfitting tendency, by making use of almost all the available lags. The constraints imposed on the parameters that assure the positivity of variance are the only reason why the algorithm is not able to fit the maximum number of lags. This phenomenon is illustrated in Table 3. Comparative results obtained with different implementations for GARCH(3,3) model on Bollerslev and Ghysels data are presented in Table 4. Unlike Matlab, the EViews solution does not respect the GARCH constraints as some of the parameters are negative. According to Alexander (2008), this would imply that the model is inappropriate for the available data. It can also be noticed that, for an adequate number of iterations I, the TA overall best estimates in terms of LLF are better than the ones found with Matlab and EViews.

Table 3: GARCH(p,q) estimators with TA optimization (I = 50,000) and LLF objective function.

maximum(p,q)	(3,3)	(5,5)	(8,8)
fitted(p,q)	(2,3)	(3,5)	(5,8)
ψ_0	-0.0037	-0.0029	-0.0038
ψ_1	0.0116	0.0142	0.0187
$\psi_{2,i}, i = 1:p$	0.1804, 0.0243	0.1930, 0.0666, 0.0051	0.1978, 0.0824, 0.0230, 0.0256, 0.0440
$\psi_{3,j}, j = 1:q$	0.2892, 9.06E-7, 0.4606	0.0106, 8.79E-4, 0.3422, 0.2870, 0.0390	2.46E-6, 5.08E-5, 0.0910, 5.71E-5, 3.97E-7, 0.1169, 0.1336, 0.2112
Conv. rate	0.955	0.944	0.926
Log-likelihood	-1096.10	-1092.49	-1088.68

The first conclusion is that LLF in (3) does not allow for an effective model selection algorithm. As the complexity of the model grows, the goodness-of-fit expressed by the value of LLF gets better, on the expense of the fitted model's power to generalize. We suggest the replacement of LLF with an objective function that satisfies the *Law of Parsimony* and penalizes additional lags, such as *Bayesian Information Criterion* (BIC):

Table 4: GARCH(3,3) estimators with different implementations with LLF as objective function.

Estimators	Implementations				
	EViews	**Matlab**	**Threshold Accepting**		
			I=1,000	I=25,000	I=100,000
ψ_0	0.00	-0.0037	-0.0053	-0.0039	-0.0035
ψ_1	0.0006	0.0113	0.0795	0.0129	0.0112
$\psi_{2,1}$	0.2182	0.1915	0.1042	0.1761	0.1825
$\psi_{2,2}$	-0.1788	0.00	0.0418	0.0459	0.0152
$\psi_{2,3}$	-0.0235	0.00	0.2359	3.63E-6	6.87E-7
$\psi_{3,1}$	1.2634	0.3880	5.44E-6	0.2185	0.3186
$\psi_{3,2}$	-0.1350	0.00	0.3280	0.0122	2.38E-6
$\psi_{3,3}$	-0.1466	0.3763	3.02E-5	0.4969	0.4399
Conv. rate	0.998	0.956	0.710	0.949	0.956
LLF	-1097.47[a]	-1096.23[a]	-1160.60	-1096.41	-1096.07

[a] The LLF values of the EViews and Matlab solutions are computed with the TA implementation.

$$BIC = -\frac{2}{T}L + \frac{n}{T}\log T,$$

where n represents the number of model parameters, L the LLF value and T the sample size.

The implementation also has to be modified according to the new optimization problem, which consists not only in parameter estimation, but also in model structure identification. Model selection deals both with finding the number of necessary lags and with identifying the holes in the structure. The modification consists mainly in adapting the heuristic for a new data structure representing a solution in the search space:

$$\Psi = \{[p_1 \dots p_{max(p)}], [q_1 \dots q_{max(q)}], \psi_0, \psi_1, [\psi_{2,1} \dots \psi_{2,max(p)}], [\psi_{3,1} \dots \psi_{3,max(q)}]\},$$

where $max(p)$ and $max(q)$ represent the maximum number of ARCH and GARCH lags, the vectors $[p_1 \dots p_{max(p)}]$ and $[q_1 \dots q_{max(q)}]$ are boolean and allow the management of structure holes, ψ_0 is the conditional mean equation parameter, ψ_1 is the constant GARCH parameter, and finally $\psi_{2,i}$ and $\psi_{3,j}$ are the ARCH (α_i) and respectively the GARCH (β_j) lag parameters.

After randomly initializing the lags, the first tendency of the heuristic is towards the reduction of the number of non-zero lags as this is the fastest way of decreasing the rate of convergence. When higher than one, this rate is responsible for explosive variance. After reaching a "stable" solution, adding of new lags with random initial values, as well as altering the current values of the existing lags are considered. In an artificial DGP experiment, it is unlikely to rediscover the true model structure because of the small and finite sample size. Minimizing the information criterion results in solutions

with fewer lags than the original seed, because of the objective function's bias towards parsimonious solutions (sometimes even entirely missing the GARCH factor). Thereby, even if the BIC values of the estimated solutions are very close to the originals', their LLF values are sometimes significantly different.

For Monte Carlo simulation purposes, we have generated 50 time series starting with the GARCH(3,3) solution below computed for the Bollerslev and Ghysels time series:

$$\Psi^{\mathrm{TR}} = [-0.003436, 0.010574, 0.186631, \mathrm{Hole}, \mathrm{Hole}, 0.364402, \mathrm{Hole}, 0.407439].$$

The heuristic was run ten times for all combinations between six different values of sample size T and five different values of the number of iterations I, summing up to $15,000$ runs. The convergence of the estimators towards their true values in dependence on the number of observations and for a fixed number of $100,000$ iterations is depicted in Figure 3.

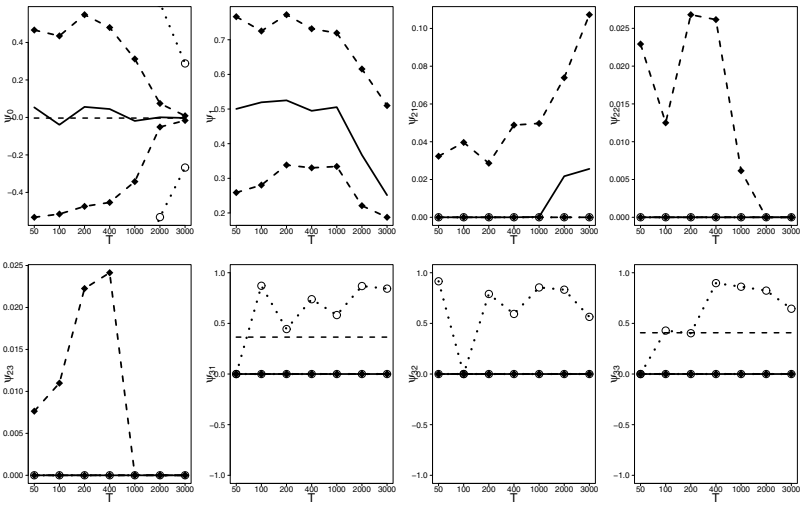


Fig. 3: The convergence of ψ_0, ψ_1, $\psi_{2,i}$, $\psi_{3,j}$ estimators, (from left to right and top to bottom) of 50 time series in relation with T: true values (horizontal dashed line), median (solid line), 25 and 75% quantiles (dashed lines with filled diamonds), minimum and maximum (dotted lines with empty circles).

As we are not able to rediscover the true model structure, it is futile to test the convergence of the estimators towards their true values. The constant parameter in (1) constitutes an exception from this statement, because the conditional mean equation has a fixed structure and influences the LLF value through ϵ_t. On the other side, the GARCH factor lags $\psi_{3,1}$—$\psi_{3,3}$ are included in less than 10% of the optimized models, while the second and third lags

of the ARCH factor, $\psi_{2,2}$ and $\psi_{2,3}$, are non-zero in less than 50%. This *zero inflation* of the estimators in (9), as well as the non-normal distribution of their values, are the reason why the OLS estimation of the linear regression described in (6) fails for most parameters, as pointed out with *'NA'* in Tables 5 and 6. The evaluation results of the convergence relation with respect to the sample size are summarized in Table 5. It can be noticed that the minimum lag size, corresponding to the biggest values of the zero-inflation indicator, is found when the heuristic number of iterations is maximum, i.e., $100,000$. When the percent of non-zero estimators is less than 50%, which is true only for ψ_0, ψ_1 and $\psi_{2,1}$, the inverse relation between $MSD^{TR,I}$ and the available number of observations holds. On the other side, the evaluation of the convergence property relating the heuristic results to the theoretical ML estimator with regard to the number of iterations is described in Table 6. For small sample sizes, the standard deviation exhibits huge volatility as the heuristic fails to converge. However, even for bigger samples, the regression analysis is weakened by the fact that as the number of iterations increases, also does the number of null model lags.

Table 5: The influence of sample size T for fixed number of iterations $I = 5000, 10000, 25000, 50000, 100000$.

	$MSD^{\mathbf{TR},I}$ **of**							
I	Ψ_0	Ψ_1	$\Psi_{2,1}$	$\Psi_{2,2}$	$\Psi_{2,3}$	$\Psi_{3,1}$	$\Psi_{3,2}$	$\Psi_{3,3}$
"Zero-inflation" of ψ_i (%)								
5000	0.00	0.00	47.60	60.37	67.07	95.27	94.73	92.10
10000	0.00	0.00	46.37	58.70	68.53	94.43	94.77	90.40
25000	0.00	0.00	43.40	60.10	68.90	95.00	94.23	88.17
50000	0.00	0.00	44.23	59.20	69.17	94.80	94.17	87.53
100000	0.00	0.00	52.29	69.54	72.86	99.06	99.11	99.06
b_p, averaged over d								
5000	-0.577	-0.634	-1.162	0.206	0.110	NA	NA	NA
10000	-0.547	-0.778	-1.216	0.218	0.093	NA	NA	NA
25000	-0.599	-0.685	-1.109	0.346	0.062	NA	NA	NA
50000	-0.580	-0.899	-1.003	0.189	0.030	NA	NA	NA
100000	-0.744	-1.911	-2.411	-0.012	-0.090	NA	NA	NA
standard deviation of b_p								
5000	0.335	1.555	0.835	0.555	0.198	NA	NA	NA
10000	0.296	1.450	0.548	0.624	0.214	NA	NA	NA
25000	0.476	1.369	0.599	0.684	0.215	NA	NA	NA
50000	0.404	1.357	0.407	0.505	0.255	NA	NA	NA
100000	0.179	1.498	3.971	0.364	0.299	NA	NA	NA
R^2, averaged over d								
5000	0.56	0.47	0.58	0.42	0.30	0.21	0.29	0.42
10000	0.54	0.48	0.58	0.40	0.34	0.17	0.26	0.43
25000	0.56	0.51	0.62	0.39	0.32	0.18	0.37	0.45
50000	0.56	0.49	0.66	0.38	0.31	0.30	0.36	0.52
100000	0.65	0.36	0.26	0.21	0.23	0.06	0.05	0.05

Table 6: The influence of the number of iterations I for fixed sample size $T = \{50, 100, 200, 400, 1000, 2000\}$.

	$MSD^{\mathrm{ML},T}$ of							
T	Ψ_0	Ψ_1	$\Psi_{2,1}$	$\Psi_{2,2}$	$\Psi_{2,3}$	$\Psi_{3,1}$	$\Psi_{3,2}$	$\Psi_{3,3}$
b_p, averaged over d								
50	-2.155	-2.032	-837.27	-1013.76	-197.97	NA	NA	NA
100	0.058	0.246	-1020	-1260	-283	NA	NA	NA
200	-0.185	0.261	-24.15	-447.85	0.222	NA	NA	NA
400	0.086	-0.094	0.043	-4.0512	-3.223	NA	NA	NA
1000	-0.251	-0.162	0.069	NA	NA	NA	NA	NA
2000	-0.404	0.123	-0.125	-0.292	NA	0.175	NA	NA
standard deviation of b_p								
50	8.248	10.40	5636.52	3025.90	5748.83	NA	NA	NA
100	4.377	1.277	7105.25	6412.81	2441.05	NA	NA	NA
200	2.092	1.229	172.00	3071.33	365.02	NA	NA	NA
400	1.984	1.087	1.781	30.918	20.84	NA	NA	NA
1000	1.376	1.145	0.796	NA	NA	NA	NA	NA
2000	1.343	0.975	0.863	1.015	NA	1.127	NA	NA
R^2, averaged over d								
50	0.29	0.27	0.33	0.34	0.39	0.18	0.04	0.05
100	0.41	0.37	0.34	0.42	0.37	0.24	0.12	0.09
200	0.37	0.27	0.37	0.32	0.35	0.15	0.21	0.19
400	0.31	0.30	0.28	0.34	0.28	0.25	0.23	0.25
1000	0.39	0.37	0.36	0.33	0.34	0.27	0.27	0.31
2000	0.32	0.39	0.38	0.36	0.38	0.36	0.29	0.42

4 Conclusions

Threshold Accepting has been successfully implemented for the parameter estimation of GARCH familiy models: GARCH(1,1), GJR-GARCH(1,1) and GARCH(p,q) with fixed number of lags. Furthermore, the convergence of the heuristic-based estimators towards their true values has been proven when the number of heuristic iterations is set up proportional to the sample size. In the case of generalized GARCH(p,q), which consists in a variable number of lags and also accepts holes in the lag structure, the standard log-likelihood function does not allow for an effective model selection algorithm and has been substituted with a Bayesian Information Criterion, replacing thus the overfitting tendency with a bias towards parsimonious solutions. Experimentally, it has been proven that the standard or normalized-standard BIC are appropriate within a heteroskedastic environment. Even so, Monte Carlo simulations and the results evaluating the relations between the mean square deviations of the estimators from their true values and from their maximum likelihood values, on one side, and the size of observations and the number of heuristic iterations, on the other side, show the complexity of the search space, as well as the convergence difficulties. The main reason is that the heuristic is able to find better fitted solutions (with fewer lags) than the original DGP seeds,

due to the finite sample sizes. However, rather than identifying the "correct" model structure, it would be of more interest to obtain good forecasts. With this goal in sight, a comparison of the relative forecasting performance exhibited by different methods is left for future research.

Acknowledgements This work is in part supported by the COST Action IC0702 and the Romanian project PN-II-RU-TE-2011-3-0242.

References

1. Alexander C (2008) *Market Risk Analysis vol. II — Practical Financial Econometrics*. J. Wiley & Sons, Chichester
2. Bollerslev T (1986) Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31:307–327
3. Bollerslev T, Ghysels E (1996) Periodic autoregressive conditional heteroscedasticity. *Journal of Business & Economic Statistics* 14(2):139–151
4. Brooks C, Burke S (2003) Information criteria for GARCH model selection. *The European Journal of Finance* 9:557–580
5. Brooks C (2008) *Introductory Econometrics for Finance*, 2nd edition. Cambridge University Press
6. Engle R (2001) GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives* 15(4):157–168
7. Gilli M, Winker P (2009) Heuristic Optimization Methods in Econometrics. In: Belsley DA, Kontoghiorghes EJ (eds.) *Handbook of Computational Econometrics*, 81–119. J. Wiley & Sons, Chichester
8. Greene W (2003) *Econometric Analysis*, 5th edition. Prentice Hall
9. Herwartz H (2004) Conditional Heteroskedasticity. In: Lutkepohl H, Kratzig M (eds) *Applied Time Series Econometrics*.
10. Maringer D (2005) *Portfolio Management with Heuristic Optimization*. Springer, Berlin
11. Maringer D, Meyer M (2008) Smooth Transition Autoregressive Models — New Approaches to the Model Selection Problem. *Studies in Nonlinear Dynamics & Econometrics* 12/1, Article 5
12. Tsay R (2005) *Analysis of Financial Time Series*, 2nd edition. Wiley Interscience
13. Winker P (2001) Optimization Heuristics in Econometrics: Applications of Threshold Accepting. J. Wiley & Sons, Chichester
14. Winker P (2006) The stochastics of threshold accepting: analysis of an application to the uniform design problem. *Proc. Computational Statistics* VI:495–503
15. Winker P, Maringer D (2009) The convergence of estimators based on heuristics: theory and application to a GARCH model. *Comput Stat* 24:533–550

Lasso–type and Heuristic Strategies in Model Selection and Forecasting

Ivan Savin[1] and Peter Winker[2]

Abstract Several approaches for subset recovery and improved forecasting accuracy have been proposed and studied. One way is to apply a regularization strategy and solve the model selection task as a continuous optimization problem. One of the most popular approaches in this research field is given by Lasso–type methods. An alternative approach is based on information criteria. In contrast to the Lasso, these methods also work well in the case of highly correlated predictors. However, this performance can be impaired by the only asymptotic consistency of the information criteria. The resulting discrete optimization problems exhibit a high computational complexity. Therefore, a heuristic optimization approach (Genetic Algorithm) is applied. The two strategies are compared by means of a Monte–Carlo simulation study together with an empirical application to leading business cycle indicators in Russia and Germany.

1 Introduction

The model selection process is crucial for the further analysis of any multiple regression model and its forecasting performance. Picking up too many regressors increases the variance of the constructed model, and taking fewer regressors than needed might result in biased and even inconsistent estimates. Both of these problems can also have negative effects on the quality of forecasts based on the models obtained through the application of these methods.

[1] DFG Research Training Program 'The Economics of Innovative Change', Friedrich Schiller University Jena and Max Planck Institute of Economics, Bachstrasse 18k Room 216, D-07743 Jena, Germany, Ivan.Savin@uni-jena.de

[2] Justus Liebig University Giessen, Licher Strasse 64, D-35394 Giessen, and Centre for European Economic Research, Mannheim, Germany,
Peter.Winker@wirtschaft.uni-giessen.de

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 165–176.
springerlink.com

During the last years, the least absolute shrinkage and selection operator (Lasso) [16] has become a very popular approach for simultaneous model selection and parameter estimation. Its main advantage is seen in obtaining both a high prediction accuracy and a parsimonious model, which is due to the regularization parameter which results in shrinking the coefficients of insignificant regressors towards zero. Hence, the resulting models concentrate on the strongest effects which tends to increase the total accuracy of the model forecast. Furthermore, the Lasso is very computationally efficient (hardly exceeding the complexity of one linear regression [3]).

However, the Lasso has some limitations. In particular, inconsistent estimates are obtained for highly correlated regressors. Numerous modifications have been suggested revising and improving the initial Lasso concept (e.g., the elastic net, the adaptive lasso), which can improve its performance under certain conditions, but do not represent a universal remedy from the limitation stated.

An alternative to the shrinkage operator is offered by model selection approaches based on information criteria (IC) which tend to provide a consistent model choice also for correlated predictors. However, even for a moderate number of predictors, these methods might result in substantial computational cost when considering a full enumeration of all alternatives. Fortunately, thanks to advances in heuristic optimization methods mimicking some evolution processes [6], there are efficient algorithms able to identify at least a good approximation to the IC's global optimum even for larger problem instances. Furthermore, IC's performance is naturally impaired by small sample sizes due to their only asymptotic consistency.

To the best of our knowledge, this study is the first[1] comparing the Lasso–type and heuristic methods both for model selection and forecasting, and contributing to the literature by demonstrating that in certain situations (e.g., if regressors in a given data set are pairwise highly correlated and for large data sets) heuristic algorithms can outperform the Lasso–type solutions.

The rest of this chapter is structured as follows. Section 2 introduces both the Lasso–type and the heuristic methods. Section 3 provides results of a Monte–Carlo analysis, and Section 4 illustrates an application to leading business cycle indicators. Finally, Section 5 concludes.

2 Model Selection Methods

The least absolute shrinkage and selection operator (Lasso), introduced by Tibshirani ([16]), is a constrained version of the ordinary least squares estimator, but has also been applied to GMM–estimators. Numerous applications

[1] An exception, however, only with regard to the comparison of the two strategies for model selection can be found in [13].

of this technique can be found in medicine, economics and other scientific fields [7] including also time series forecasting (see, among others, [1]).

Consider the model selection problem for the following regression function:

$$y = \alpha + X^{opt}\beta + \varepsilon, \tag{1}$$

where α is an n-vector with all elements equal, X is an $n \times k$ matrix of k regressors and their values for n observations, β is a $k \times 1$ vector of their coefficients and ε is an $n \times 1$ vector of residuals. In (1) X^{opt} refers either to the 'true' model in a Monte–Carlo simulation set–up or to an optimal approximation to the unknown real data generating process. Standardizing the predictors so that they have mean 0 and standard deviation equaling 1, and the response having mean 0, one can omit α without loss of generality.

2.1 Lasso–type Strategies

For (1) the Lasso objective function can be presented as follows:

$$\widehat{\beta}_{Lasso} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda \| \widehat{\beta} \|_1 \right]. \tag{2}$$

While the first term in the right part of equation (2) measures the fit of the model by the residual sum of squares (RSS), the second one with $\lambda > 0$ is the shrinkage applied to the sum of the absolute values of the coefficients. Hence, the Lasso can be referred to as a special case of the Bridge regression approach [4] imposing an upper bound on the L^q-norm of the parameters $(0 < q < \infty)$ with $q = 1$:

$$\| \widehat{\beta} \|_q = \left[\sum_{j=1}^{k} |\beta_j|^q \right]^{1/q}. \tag{3}$$

There are different approaches to solve (2) including quadratic programming, coordinate–wise optimization and gradient projection (see, e.g., [5]). For the sake of brevity we do not discuss any of those methods, so that the interested reader is advised to consult the literature. In this study we use a modification of the *LARS* algorithm suggested by Efron et al. ([3]) and popularized among practitioners.[2] The algorithm provides a piecewise–linear solution path in the tuning parameter $\lambda \in [0, \infty)$ with all $\widehat{\beta}$'s set to zero at $\lambda = \infty$ and equal to the OLS estimate at $\lambda = 0$. To select a single solution, λ is chosen by tenfold cross–validation minimizing the prediction error (PE) of the model.

[2] Related code is available at http://www.stanford.edu/~hastie/Papers/LARS for R and http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897 for Matlab.

Setting $\lambda > 0$ by cross–validation one insures the Lasso solution to have a parsimony property, i.e. only a subset of resulting predictors in (2) has non–zero coefficients. This feature of the Lasso might increase the total accuracy of the model forecast and improves the interpretability of the selected model.

However, the Lasso has substantial limitations. First, it cannot identify all 'true' predictors in a data set with pairwise highly correlated regressors [21]. The latter can be referred to as the 'irrepresentable condition' [19, p. 2544]. Thus, Lasso is consistent in low correlation settings only, when

$$\max_{j>r} \| cov(X_j, X^{true}) cov(X^{true})^{-1} \|_1 < 1, \tag{4}$$

while in presence of high correlations between 'true' and irrelevant variables, the Lasso cannot recover the correct sparsity pattern ($\widehat{\beta}_{Lasso} \nrightarrow \beta^{true}$).

However, as Meinshausen and Yu ([10]) show, even failing to discover the correct sparsity pattern (when (4) does not hold), the Lasso can provide good approximations of the 'true' model for large sample sizes ($\| \beta - \widehat{\beta}_{Lasso} \|_2 \to 0$ as $n \to \infty$). In other words, Lasso selects 'true' variables with high probability and irrelevant ones have only marginal coefficients (L^2-norm consistency).

Second, Lasso is inconsistent for $k \gg n$ (underdetermined linear system), where (2) can identify not more than $n - 1$ (standardized) predictors [3].

Lasso Modifications

Many proposals have been made on how to improve the Lasso concept. Due to space restrictions, we concentrate only on two such modifications. For a more complete overview, the interested reader is referred, e.g., to [7] and [5].

We consider two extensions of the Lasso: the elastic net (EN) using a combination of the Lasso (λ_1) and the ridge regression (λ_2) penalty [21]:

$$\widehat{\beta}_{EN} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda_1 \| \widehat{\beta} \|_1 + \lambda_2 \| \widehat{\beta} \|_2^2 \right], \tag{5}$$

and the adaptive Lasso (aLasso) applying different amounts of shrinkage for each regression coefficient [20].[3]

$$\widehat{\beta}_{aLasso} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda \sum_{j=1}^{k} \widehat{\omega}_j |\beta_j| \right]. \tag{6}$$

Thus, the selected extensions are particularly designed to deal with the limitations stated and operate in a continuous space remaining computationally efficient. Furthermore, in line with [2] we also perform unregularized restricted estimation (i.e. OLS estimation on the selected set of regressors)

[3] For more details the reader is referred to the literature. See also [13].

for all Lasso–type methods tested to alleviate a potential bias that results from the regularized estimation ($\lambda > 0$).

2.2 Heuristic Optimization Methods

Alternatively, information criteria (IC) can be used to identify X^{opt} in (1). IC rank different models according to their fitness, while penalizing model complexity. Hence, they can be interpreted as a L^0-constraint, penalizing not the coefficients' values, but only their number:

$$\widehat{\beta}_{IC} = \arg\min_{\beta} \left[\| y - X\widehat{\beta} \|_2^2 + \lambda \, \| \widehat{\beta} \|_0 \right]. \tag{7}$$

IC have become a standard instrument in model selection ranging from lag order selection in multivariate linear and nonlinear autoregression models to selection between rival nonnested models [18]. In this study, the Bayesian IC (BIC) and the Hannan-Quinn IC (HQIC) are employed. For infinitely large sample sizes these IC are consistent model selection instruments and, as noted by Hastie et al. ([19, p. 2553]), the solution of (7) remains consistent even for data sets with correlated regressors.

Given that the search space of candidate models in (7) is discrete, standard gradient methods cannot be applied. Also the full enumeration of all possible solutions is only feasible for a moderate k. Consequently, in the last decade many studies have been devoted to the problem in (1): sequential bottom–up (top–down) inclusion (deletion) of individual regressors [12, 8]; usage of certain prior probabilities shrinking the parameter search space and resulting in model averaging [9]. However, these methods investigate only a specific fraction of all submodels, whereas there is no guarantee to find the 'true' model in this way.

In order to tackle the highly complex integer optimization problem, one can take advantage of optimization heuristics that mimic natural evolution processes. These methods are called 'heuristic' or 'meta-heuristics' because of their stochastic nature that helps them to converge to a model which at least represents a good approximation to the IC optimum. For an overview of these optimization techniques see [6]. In [15] a similar subset selection problem was handled by two algorithms: Threshold Accepting and Genetic Algorithms (GA). Since GA provided slightly better results in terms of both CPU time and solution quality, only GA are considered in the following.

GA are population–based heuristics that investigate the search space in many directions simultaneously, performing jumps in the search space by means of crossover and mutation mechanisms. Thereby, the probability of getting stuck in a local optimum is reduced. The members in the population are represented as bit strings of ones and zeros corresponding to the predictor variables included and not included in the candidate model. In each

generation GA replace parts of a population with new solutions aimed to be better for a given problem. The GA algorithm implemented is very similar to the one in [15].[4] The only difference is that 1000 generations are found to be sufficient in this study for GA to converge.

3 Monte–Carlo Study

The goal of this section is to determine in what set–ups which of the two strategies, Lasso–type methods (Lasso, EN, aLasso) or GA tined by IC, provide superior results (in terms of correctly recovered subsets, forecasting and estimation accuracy) and what is the corresponding CPU-time required.

Data Generating Process

To this end, different artificial data sets are generated varying the sample size (n) from 100 (frequent in macroeconomics) to 1000 (which is mostly available only in finance and natural sciences) and fixing the number of potential regressors to 50. First, we generate 4 predictors with a joint Gaussian distribution and covariance matrix Σ. We choose either $\Sigma_{i,j} = 0.5^{|i-j|}$ or $0.75^{|i-j|}$ with $1 \leq i, j \leq k$, corresponding to a 'low' and 'high' correlation setting, respectively. Second, the data matrix consisting of lags 1 to 10 of these predictors is formed (X^{mc}). Third, we select a small number of elements $k^{true} = 5$ of the coefficient vector β^{mc}, which are set to non–zero values.[5] These non–zero coefficient values are randomly distributed between -1 and 1, and divided by the respectively chosen lag order so that lags of higher order are (on average) assigned with smaller coefficients.[6] Fourth, the initial value of the response variable (y_0^{mc}) is set to zero, and based on β_j^{mc}, one recursively generates y_t^{mc} and adds an i.i.d. normal random error term.[7]

$$y_t^{mc} = \sum_{i=1}^{10} \beta_{0,i}^{mc} y_{t-i}^{mc} + \sum_{j=1}^{4} \sum_{i=1}^{10} \beta_{j,i}^{mc} x_{j,t-i}^{mc} + \varepsilon_t, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2). \tag{8}$$

In (8) one chooses σ_ε such that the corresponding noise–to–signal ratio (NSR, for details see [4, p. 125]) equals either 1/5 ('low noise') or 2 ('high noise'). Obviously, (8) represents an Autoregressive Distributed Lag model with 10

[4] Thus, a population of 500 solutions, the uniform crossover mechanism and a mutation operator applied to 5 randomly chosen genes with 50% probability are employed.

[5] One ensures that one lag of each variable (including the dependent one) is included.

[6] This appears reasonable since in empirical studies lags of lower order are found to be more important.

[7] Finally, the first 11 observations in y^{mc} and X^{mc} are discarded.

lags for both, one dependent and four explanatory variables, where no current values of the explanatory variables are involved. Thus, for a general ADL(p_1, p_2, p_3) we consider ADL(10,4,10).

Simulation Results

The quality of the results in terms of model identification is assessed by the True Positive Rate (TPR) and the False Negative Rate (FNR)[8], whereas mean–squared error (MSE $= E[(\widehat{\beta} - \beta^{mc})' \Sigma (\widehat{\beta} - \beta^{mc})]$) is used as a measure of the estimation accuracy.[9] For this purpose, 90% of the observations are used as a training set. The CPU time corresponding to a single restart using Matlab 7.11 on a Pentium IV 3.3 GHz is reported.[10]

Furthermore, the remaining 10% of observations are left for an out–of–sample forecast, where root mean–squared forecast error, and its standard deviation computed over 50 replications (in parentheses),

$$RMSFE = \sqrt{\frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} (y_t^{mc} - \widehat{y}_t)^2}, \tag{9}$$

is used to assess the forecast quality. Thereby, T_1 and T_2 indicating the first and the last period of the forecasting period.

Simulation results obtained for different set–ups are reported in Table 1.[11] For medium–sized samples ($n = 500$) heuristics clearly outperform Lasso–type methods in subset recovery and estimation accuracy,[12] which eventually results in a better forecasting performance. However, the difference in RMSFEs is not that large. Among the Lasso methods considered, aLasso provides in general superior results, and this dominance holds for different correlation and noise settings. For 'high correlation' some marginal improvements as compared to classical Lasso are obtained via EN, which is due to the more robust ridge penalties.

It can be observed that heuristics improve in performance relative to the Lasso methods in low noise settings and for larger sample sizes. The former is due to a more restrictive selection performed by the shrinkage strategies, which for 'low noise' translates in substantially more type II errors, i.e., ignor-

[8] TPR is the percentage of 'true' regressors from all variables selected and FNR is the portion of rejected 'true' regressors among correctly selected and correctly rejected ones.

[9] Standard deviations computed over 50 replications are given in parenthes in Table 1. Unregularized restricted estimations for Lasso–type methods are reported as MSE_2.

[10] For each method, averages over 50 replications of the procedure are reported.

[11] Due to space constraints, here we report only results for the BIC, but qualitatively similar findings based on HQIC are available on request.

[12] Even accounting for MSE_2 an improvement for all scenarios is depicted in Table 1.

ing too many relevant predictors, whereas the latter results from the asymptotic consistency of IC allowing to identify the correct sparsity pattern. In contrast, for $n = 100$ the difference between shrinkage and heuristic methods becomes less evident. Furthermore, for 'high noise' and small n Lasso–type strategies indisputably beat heuristics both in estimation and forecasting.

Table 1 Monte–Carlo simulation results

			Lasso	EN	aLasso	BIC	Lasso	EN	aLasso	BIC
			Low correlation				High correlation			
Results for $n = 100$	Low noise	TPR	72.7%	71.2%	70.9%	56.0%	73.9%	73.5%	60.7%	53.8%
		FNR	2.6%	2.5%	2.3%	1.0%	3.0%	2.9%	3.0%	1.6%
		MSE	.0212	.0209	.0102	.0035	.0235	.0233	.0104	.0076
			(.0413)	(.0408)	(.0160)	(.0033)	(.0431)	(.0422)	(.0140)	(.0166)
		CPU	.4s	1.0s	1.0s	32s	.3s	.7s	1.0s	31s
		MSE_2	.0166	.0165	.0076		.0189	.0188	.0083	
			(.0374)	(.0373)	(.0135)		(.0374)	(.0372)	(.0143)	
		RMSFE	.0114	.0113	.0112	.0113	.0119	.0117	.0126	.0122
			(.0053)	(.0051)	(.0048)	(.0039)	(.0049)	(.0049)	(.0052)	(.0053)
	High noise	TPR	79.3%	79.3%	62.7%	36.6%	71.2%	70.2%	50.0%	31.9%
		FNR	7.1%	7.1%	7.1%	6.1%	7.2%	7.3%	7.0%	6.6%
		MSE	.0339	.0336	.0247	.0529	.0363	.0359	.0286	.0521
			(.0718)	(.0719)	(.0447)	(.0542)	(.0856)	(.0857)	(.0527)	(.0518)
		CPU	.3s	.7s	1.1s	28s	.3s	.7s	.7s	29s
		MSE_2	.0234	.0234	.0224		.0338	.0338	.0234	
			(.0451)	(.0451)	(.0413)		(.0844)	(.0844)	(.0421)	
		RMSFE	.1078	.1078	.1107	.1191	.1088	.1089	.1123	.1234
			(.0402)	(.0402)	(.0424)	(.0496)	(.0415)	(.0413)	(.0449)	(.0547)
Results for $n = 500$	Low noise	TPR	68.5%	71.5%	68.3%	84.5%	66.1%	67.0%	74.6%	87.1%
		FNR	2.1%	2.1%	1.6%	.7%	2.3%	2.1%	1.6%	.7%
		MSE	.0187	.0187	.0069	3.0×10^{-4}	.0209	.0208	.0063	2.3×10^{-4}
			(.0299)	(.0299)	(.0158)	(4.7×10^{-4})	(.0376)	(.0376)	(.0196)	(3.1×10^{-4})
		CPU	.3s	.8s	.8s	88s	.3s	1.0s	.9s	93s
		MSE_2	.0120	.0120	.0023		.0154	.0151	.0042	
			(.0235)	(.0235)	(.0052)		(.0328)	(.0329)	(.0182)	
		RMSFE	.0102	.0102	.0102	.0098	.0107	.0107	.0107	.0103
			(.0041)	(.0041)	(.0042)	(.0041)	(.0043)	(.0043)	(.0046)	(.0041)
	High noise	TPR	95.6%	95.6%	75.4%	81.8%	91.7%	91.7%	78.4%	79.7%
		FNR	6.3%	6.3%	5.0%	3.5%	6.4%	6.4%	5.7%	4.1%
		MSE	.0269	.0266	.0214	.0054	.0270	.0266	.0225	.0060
			(.0420)	(.0413)	(.0488)	(.0056)	(.0426)	(.0415)	(.0432)	(.0063)
		CPU	.3s	.8s	.8s	78s	.3s	.8s	.8s	81s
		MSE_2	.0166	.0166	.0087		.0167	.0167	.0108	
			(.0276)	(.0276)	(.0105)		(.0277)	(.0277)	(.0191)	
		RMSFE	.0985	.0985	.0962	.0957	.1007	.1006	.0992	.0981
			(.0407)	(.0407)	(.0392)	(.0397)	(.0421)	(.0421)	(.0407)	(.0413)
Results for $n = 1000$	Low noise	TPR	54.2%	54.3%	80.8%	90.8%	60.1%	60.1%	72.2%	88.7%
		FNR	1.4%	1.4%	.9%	.4%	1.7%	1.7%	.9%	.3%
		MSE	.0107	.0107	.0041	6.9×10^{-5}	.0122	.0122	.0030	1.3×10^{-4}
			(.0181)	(.0181)	(.0077)	(1.1×10^{-4})	(.0171)	(.0172)	(.0048)	(2.6×10^{-4})
		CPU	.4s	1.0s	1.7s	157s	.5s	1.5s	1.8s	156s
		MSE_2	.0044	.0044	6.7×10^{-4}		.0080	.0080	.0010	
			(.0097)	(.0097)	(.0019)		(.0132)	(.0132)	(.0033)	
		RMSFE	.0103	.0103	.0102	.0101	.0095	.0095	.0098	.0092
			(.0038)	(.0038)	(.0037)	(.0037)	(.0040)	(.0040)	(.0067)	(.0039)
	High noise	TPR	93.7%	93.9%	75.7%	85.9%	92.9%	93.0%	76.3%	83.2%
		FNR	5.2%	5.2%	3.6%	2.3%	5.3%	5.4%	4.4%	2.7%
		MSE	.0199	.0199	.0089	.0029	.0199	.0198	.0153	.0030
			(.0247)	(.0247)	(.0115)	(.0042)	(.0245)	(.0245)	(.0248)	(.0043)
		CPU	.5s	1.5s	1.4s	147s	.5s	1.4s	1.3s	153s
		MSE_2	.0132	.0133	.0063		.0133	.0132	.0086	
			(.0178)	(.0178)	(.0109)		(.0178)	(.0178)	(.0145)	
		RMSFE	.0925	.0927	.0912	.0902	.0951	.0949	.0944	.0926
			(.0378)	(.0386)	(.0375)	(.0372)	(.0403)	(.0397)	(.0403)	(.0395)

4 Application to Leading Business Cycle Indicators in Germany and Russia

Being particularly interested in the usefulness of the strategies from the forecasting point of view, we also show their application to real economic data.

Leading indicators (LI) are nowadays a standard tool for the analysis and forecasting of business cycles due to the publication delay of data on real production. While for Germany (as for other industrial countries) there is a large body of empirical evidence that models forecasting industrial production (IP), which include LI, outperform forecasts of univariate time series models [17, 11], there is less such evidence for developing countries.

For the empirical application we use two LI (business expectations and business climate) and IP for Germany and Russia for the period 02/1999–09/2009. More information on the properties of the data can be found in [14]. Important is that the German LI are seasonally adjusted, while for Russia they are not. Furthermore, a potential structural break in Russian data must be accounted. Hence, we consider the IP indices for both countries also as unadjusted and introduce seasonal and shift dummies, and their interaction terms (for details see [14]) to account for these data features.

Similar to Section 3, ADL models (augmented with seasonal and shift dummies) are our modelling framework to identify predictors and construct forecasts, while an AR(2) process serves as benchmark. The latter is found to be a hard competitor in business cycle forecasting for small data sets [14].

We only employ 1–step–ahead forecasts of IP growth rates (log differences) for periods of one and two years length between 11/2006 and 09/2009 (this also allows one to consider the forecasting performance both prior and during the crisis) and increasing estimation windows (IEW). Furthermore, in contrast to [14], we allow for lags from both LI and both countries to be included in (9) for each IP, so that a selection out of 65 predictors (5 variables and 13 lags) has to be made. As a result, two data sets with highly correlated potential predictors are generated (see Figure 1).[13]

Results for the two model selection strategies are exhibited in Table 2. As one can see, for the final forecasting period the shrinkage strategies mostly dominate the benchmark for both 12– and 24–month period forecasts, while heuristics fail do so. The main reason for this performance is seen to be the small estimation sample available: there are merely 128 observations for estimation and forecasting in total, which is most comparable with the upper panel in Table 1. Furthermore, since the IP growth rates can be to a large extent ($R^2 \approx 70 - 80\%$) explained by the set of lags selected (together with the seasonal and shift dummies), which corresponds to the low noise setting, EN and Lasso outperform aLasso. Finally, since particularly for small noise and high correlation among predictors in small samples Lasso–type methods

[13] The main diagonal in the correlation matrix is removed.

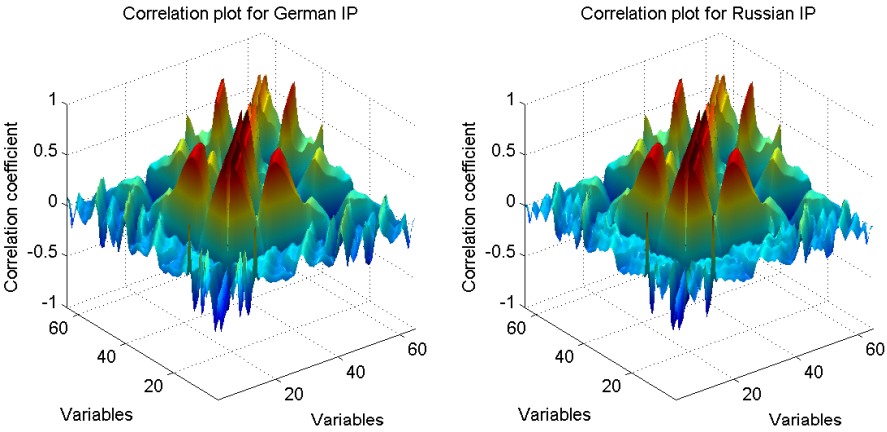


Fig. 1 Pairwise correlations in the empirical data sets

Table 2 Forecasting performance of the ADL models

Model specification			Germany	Russia	Germany	Russia
			10/2007–09/2009		10/2008–09/2009	
RMSFE in relation to AR(2)	Lasso		0.9064	0.8296	0.7479	0.7800
	EN		0.9469	0.8296	0.7479	0.7748
	aLasso		0.9849	0.9744	1.1355	0.8422
	Genetic Algorithms	BIC	1.1763	0.9629	1.1832	0.9322
		HQIC	1.1558	1.1191	1.2473	1.1169

provide some better forecasts than heuristics, the advantage of the shrinkage methods could be expected from the Monte–Carlo results.

We also consider the performance of the two strategies over a set of forecasting periods shifting by one month (rolling windows). The results are provided in Figure 2 (upper panel for 12– and lower for 24–month forecasts).

5 Conclusions

Since the correct dynamic specification of time series models is often unknown, the use of model selection strategies is required. We consider two classes of model selection approaches, one based on shrinkage estimators such as Lasso and the other one – a subset selection method making use of optimization heuristics, to solve the corresponding highly complex discrete optimization problem.

A Monte–Carlo simulation is used to assess the merits of the different methods in the context of univariate autoregressive distributed lag models.

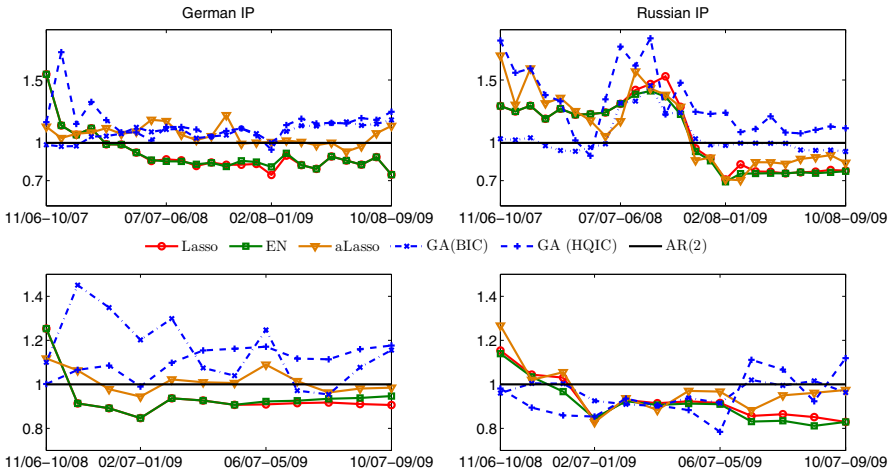


Fig. 2 Forecast accuracy in relation to AR(2) with IEW (RMSFE in relation to AR(2))

The simulation setting is chosen to mimic realistic situations found in the framework of forecasting business cycles. In particular, the number of available observations is often small compared to the number of potentially relevant predictors. Due to the high persistence in many economic variables, different lags of these predictors might be highly correlated rendering the model selection problem more challenging. The results from the Monte–Carlo simulation suggest that the use of information criteria in the subset selection approach is impaired by the small number of observations, while the shrinkage estimators still perform remarkably well despite of the high correlation of potential predictors.

Furthermore, we consider to what extent a proper model selection might help to improve forecasts of business cycle indicators for Russia and Germany. While the improvements compared to a simple autoregressive process are small in all settings, we find again slight advantages of the shrinkage estimators.

Based on these findings, several questions emerge naturally which we will consider in future research. In particular, we will test whether larger sample sizes improve the relative performance of information criteria based selection as these criteria are asymptotically consistent. Furthermore, we will study a situation with a larger number of relevant regressors in the model. Finally, further real applications will be studied to learn about performance gains to be expected when moving away from simplistic univariate time series models.

Acknowledgements Financial support from the German Science Foundation (DFG RTG 1411) is gratefully acknowledged.

References

1. Bai J, Ng S (2008) Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146(2):304–317
2. Candès EJ, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* 35(6):2313–2351
3. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Annals of Statistics* 32:407–489
4. Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. *Technometrics* 35(2):109–135
5. Gasso G, Rakotomamonjy A, Canu S (2009) Recovering sparse signals with a certain family of non-convex penalties and DC programming. *IEEE Trans. on Signal Processing* 57(12):4686–4698
6. Gilli M, Winker P (2009) Heuristic optimization methods in econometrics. In: Belsley D, Kontoghiorghes E (eds.) *Handbook of Computational Econometrics*, 81–119. Wiley, Chichester
7. Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York
8. Hendry DF, Krolzig HM (2005) The properties of automatic 'GETS' modelling. *The Economic Journal* 115(502):C32–C61
9. Kapetanios G, Labhard V, Price S (2008) Forecasting using Bayesian and information-theoretic model averaging: An application to U.K. inflation. *Journal of Business & Economic Statistics* 26(1):33–41
10. Meinshausen N, Yu B (2008) Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* 37(1):246–270
11. Ozyldrim A, Schaitkin B, Zarnowitz V (2010) Business cycles in the Euro area defined with coincident economic indicators and predicted with leading economic indicators. *Journal of Forecasting* 29(1–2):6–28
12. Perez-Amaral T, Gallo GM, White H (2003) A flexible tool for model building: The relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics* 65(1):821–838
13. Savin, I (2010) A comparative study of the lasso–type and heuristic model selection methods. *COMISEF Working Paper Series* 42
14. Savin I, Winker P (forthcoming) Heuristic optimization methods for dynamic panel data model selection. Application on the Russian innovative performance. *Computational Economics*
15. Savin I, Winker P (2011) Heuristic model selection for leading indicators in Russia and Germany. *MAGKS Joint Discussion Paper Series in Economics* 01–2011
16. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* 58(1):267–288
17. Vogt G (2007) The forecasting performance of ifo-indicators under realtime conditions. *Journal of Economics and Statistics* 227(1):87–101
18. Winker P (1995) Identification of multivariate AR-models by threshold accepting. *Computational Statistics & Data Analysis* 20(3):295–307
19. Zhao P, Yu B (2006) On model selection consistency of lasso. *Journal of Machine Learning Research* 7:2541–2563
20. Zou H (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101:1418–1429
21. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* 67(2):301–320

Streaming-Data Selection for Gaussian-Process Modelling

Dejan Petelin[1] and Juš Kocijan[1,2]

Abstract The Gaussian-process (GP) model is an example of a probabilistic, non-parametric model with uncertainty predictions. It can be used for the modelling of complex, non-linear systems and also for the identification of dynamic systems. The output of the GP model is a normal distribution, expressed in terms of the mean and the variance. One of the noticeable drawbacks of a system identification with GP models is the computation time necessary for the modelling. The modelling procedure involves the inverse of the covariance matrix, which has the dimension as large as the length of the input samples vector. The computation time for this inverse, regardless of the use of an efficient algorithm, is increasing with the third power of the number of input data. In this chapter we propose a method for the sequential selection of streaming data so that the size of the active set remains constrained. Furthermore, for better adjustment of the model to the system the hyperparameter values are optimised as well. The viability of the proposed method is tested on data obtained from two, nonlinear, dynamic systems.

1 Introduction

Gaussian-process (GP) models [11] form a new, emerging ,complementary method for non-linear, dynamic, system identification. The GP model is a probabilistic, non-parametric, black-box model. It differs from most other frequently used black-box identification approaches in that it does not approximate the modelled system by fitting the parameters of the selected basis functions, but rather it searches for the relationship among the measured data. GP models are closely related to approaches such as support vector machines and, in particular, relevance vector machines. Because the GP model

[1] Jožef Stefan Institute, Ljubljana, Slovenia, {dejan.petelin,jus.kocijan}@ijs.si
[2] University of Nova Gorica, Nova Gorica, Slovenia

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 177–190.
springerlink.com

is a Bayesian model, its output is a normal distribution, expressed in terms of the mean and the variance. The mean value represents the most likely output, and the variance can be viewed as a measure of its confidence. The obtained variance, which depends on the amount of available identification data, is important information that distinguishes the GP models from other non-Bayesian methods. GP models can be used for model identification when the data are noisy and when there are outliers or gaps in the input data. Another useful attribute of GP models is the possibility to include various kinds of prior knowledge into the model, e.g., local models, static characteristics, etc.

A noticeable drawback of any system identification with GP models is the computation time necessary for the modelling. Regression based on GP models involves several matrix computations in which the load increases with the third power of the number of input data, such as the matrix inversion and the calculation of the log-determinant of the used covariance matrix. This computational greed restricts the amount of training data to at most a few thousand cases. To overcome the computational-limitation issues and also to make use of the method for large-scale dataset applications, numerous authors have suggested various sparse approximations [2, 3]. A common property of all sparse approximate methods is that they try to retain the bulk of the information contained in the full training dataset, but reduce the size of the resultant covariance matrix so as to facilitate a less computationally demanding implementation of the GP model. Special kinds of sparse approximate methods are the *on-line* modelling methods, e.g., Sparse On-line Gaussian Processes [4], Fast Forward Selection to Speed Up Sparse Gaussian Process Regression [12] and Online Sparse Matrix Gaussian Process Regression and Vision Applications [10]. All these on-line methods try to incorporate all the information about the data by projecting to a reduced covariance matrix.

The problem we focus on in this chapter is the on-line selection of particular data with rich information content from streaming data to be used afterwards for the modelling of a dynamic system. This kind of modelling is of interest, for example, in the case of changing a system's dynamics with a change of the system's operating conditions. For this purpose the current sparse methods for GP models are not appropriate as they need all the training data available at once or cannot adjust the hyperparameter values in on-line mode. Therefore, we propose a streaming-data selection method for GP modelling that sequentially gains the bulk of the information contained in the streaming data and adjusts with the incoming data. To keep the subset of the most informative data small enough to process the current data before new data arrive, the maximum length of the subset is fixed. The information that each piece of data in the subset contains regarding the other data in the subset is estimated with the log-marginal likelihood. To test the viability of the proposed method it is compared to the full GP model trained on the entire training dataset and the Sparse On-line Gaussian Processes method (OGP) [4].

The outline of the chapter is as follows. Section 2 introduces GP models. Section 3 describes the proposed method for streaming-data selection for modelling with Gaussian processes. The results of the experiments are given in Section 4. Section 5 concludes the paper with a summary of the work and indicates the direction for future work.

2 Modelling with Gaussian Processes

A GP model is a flexible, probabilistic, non-parametric model for the prediction of output-variable distributions. Its properties and application potentials are reviewed in [11].

A Gaussian process is a collection of random variables that have a joint multivariate Gaussian distribution (Fig. 1). Assuming a relationship of the form $y = f(\mathbf{x})$ between the input $\mathbf{x}$ and the output y, we have $y_1, \ldots, y_N \sim \mathcal{N}(0, \mathbf{\Sigma})$, where $\Sigma_{pq} = \mathrm{Cov}(y_p, y_q) = C(\mathbf{x}_p, \mathbf{x}_q)$ gives the covariance between the output points corresponding to the input points $\mathbf{x}_p$ and $\mathbf{x}_q$. Thus, the mean $\mu(\mathbf{x})$ and the covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ fully specify the Gaussian process.

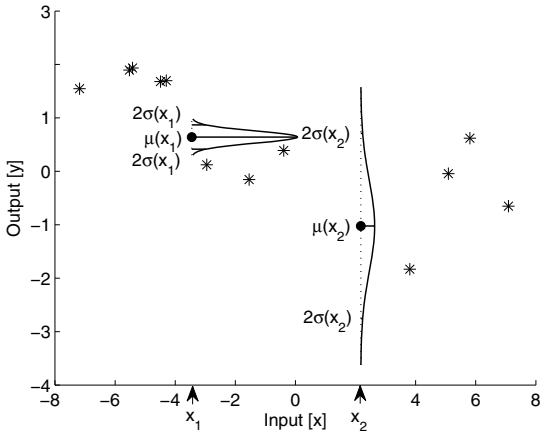


Fig. 1 Modelling with GP: Gaussian distribution of predictions at new points x_1 and x_2, conditioned on the training points (*).

The value of the covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ expresses the correlation between the individual outputs $f(\mathbf{x}_p)$ and $f(\mathbf{x}_q)$ with respect to the inputs $\mathbf{x}_p$ and $\mathbf{x}_q$. Note that the covariance function $C(\cdot, \cdot)$ can be any function that generates a positive semi-definite covariance matrix. It is usually composed of two parts,

$$C(\mathbf{x}_p, \mathbf{x}_q) = C_{\mathrm{f}}(\mathbf{x}_p, \mathbf{x}_q) + C_{\mathrm{n}}(\mathbf{x}_p, \mathbf{x}_q), \tag{1}$$

where C_{f} represents the functional part and describes the unknown system we are modelling, and C_{n} represents the noise part and describes the model of the noise.

For the noise part it is most common to use the constant covariance function, presuming white noise. The choice of the covariance function for the functional part also depends on the stationarity of the process. Assuming stationary data the most commonly used covariance function is the square exponential covariance function. The composite covariance function is therefore

$$C(\mathbf{x}_p, \mathbf{x}_q) = v_1 \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d (x_{dp} - x_{dq})^2\right] + \delta_{pq} v_0, \tag{2}$$

where w_d, v_1 and v_0 are the 'hyperparameters' of the covariance function, D is the input dimension, and $\delta_{pq} = 1$ if $p = q$ and 0 otherwise. In contrast, assuming non-stationary data the polynomial or its special case, the linear covariance function, can be used. Other forms and combinations of covariance functions suitable for various applications can be found in [11]. The hyperparameters can be written as a vector $\boldsymbol{\Theta} = [w_1, \ldots, w_D, v_1, v_0]^T$. The parameters w_d indicate the importance of the individual inputs: if w_d is zero or near zero, it means the inputs in dimension d contain little information and could possibly be neglected.

To accurately reflect the correlations present in the training data, the hyperparameters of the covariance function need to be optimised. Due to the probabilistic nature of the GP models, the common model optimisation approach, where model parameters and possibly also the model structure are optimised through the minimization of a cost function defined in terms of model error (e.g., mean square error), is not readily applicable. A probabilistic approach to the optimisation of the model is more appropriate. Actually, instead of minimizing the model error, the probability of the model is maximised.

GP models can be easily utilized for a regression calculation. Consider a matrix $\mathbf{X}$ of N D-dimensional input vectors where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$ and a vector of the output data $\mathbf{y} = [y_1, y_2, \ldots, y_N]$. Based on the data $(\mathbf{X}, \mathbf{y})$, and given a new input vector $\mathbf{x}^*$, we wish to find the predictive distribution of the corresponding output y^*. Based on the training set $\mathbf{X}$, a covariance matrix $\mathbf{K}$ of size $N \times N$ is determined. The overall problem of learning unknown parameters from data corresponds to the predictive distribution $p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*)$ of the new target y, given the training data $(\mathbf{y}, \mathbf{X})$ and a new input $\mathbf{x}^*$. In order to calculate this posterior distribution, a prior distribution over the hyperparameters $p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X})$ can first be defined, followed by the integration of the model over the hyperparameters

$$p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*) = \int p(y^*|\boldsymbol{\Theta}, \mathbf{y}, \mathbf{X}, \mathbf{x}^*) p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X}) d\boldsymbol{\Theta}. \tag{3}$$

The computation of such integrals can be difficult due to the intractable nature of the non-linear functions. A solution to the problem of intractable integrals is to adopt numerical integration methods such as the Monte-Carlo approach. Unfortunately, significant computational efforts may be required to achieve a sufficiently accurate approximation.

In addition to the Monte-Carlo approach, another standard and general practice for estimating hyperparameters is the maximum-likelihood estimation, i.e., to minimise the following negative log-likelihood function:

$$\mathcal{L}(\boldsymbol{\Theta}) = -\frac{1}{2}\log(\mid \mathbf{K} \mid) - \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi) \tag{4}$$

As the likelihood is, in general, non-linear and multi-modal, efficient optimisation routines usually entail the gradient information. The computation of the derivative of $\mathcal{L}$ with respect to each of the parameters is as follows

$$\frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\theta_i} = -\frac{1}{2}\text{trace}\left(\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_i}\right) + \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_i}\mathbf{K}^{-1}\mathbf{y}. \tag{5}$$

For performing a regression, the availability of the training set $\mathbf{X}$ and the corresponding output set $\mathbf{y}$ is assumed. Based on the training set $\mathbf{X}$, a covariance matrix $\mathbf{K}$ of size $N \times N$ is determined. The aim is to find the distribution of the corresponding output y^* for some new input vector $\mathbf{x}^* = [x_1(N+1), x_2(N+1), \ldots, x_D(N+1)]^T$.

For the collection of random variables $[y_1, \ldots, y_N, y^*]$ we can write:

$$[\mathbf{y}, y^*] \sim \mathcal{N}(0, \mathbf{K}^*) \tag{6}$$

with the covariance matrix

$$\mathbf{K}^* = \left[\begin{array}{c|c} \mathbf{K} & \mathbf{k}(\mathbf{x}^*) \\ \hline \mathbf{k}^T(\mathbf{x}^*) & \kappa(\mathbf{x}^*) \end{array}\right] \tag{7}$$

where $\mathbf{y} = [y_1, \ldots, y_N]$ is a $1 \times N$ vector of training targets. The predictive distribution of the output for a new test input has a normal probability distribution with a mean and variance

$$\mu(y^*) = \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\mathbf{y}, \tag{8}$$

$$\sigma^2(y^*) = \kappa(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}^*), \tag{9}$$

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}_1, \mathbf{x}^*), \ldots, C(\mathbf{x}_N, \mathbf{x}^*)]^T$ is the $N \times 1$ vector of covariances between the test and training cases, and $\kappa(x^*) = C(\mathbf{x}^*, \mathbf{x}^*)$ is the covariance between the test input itself.

The obtained model, in addition to the mean value, provides information about the confidence in the prediction by the variance. Usually, the confidence of the prediction is depicted with a 2σ interval, which is an about 95% confidence interval. This confidence region can be seen in the example in Fig. 2 as a grey band. It highlights the areas of the input space where the prediction quality is poor, due to the lack of data or noisy data, by indicating a wider confidence band around the predicted mean.

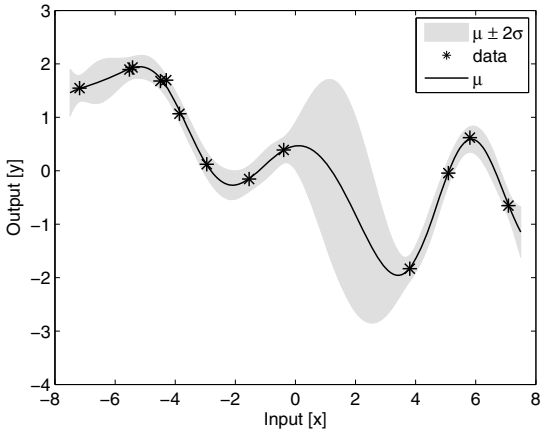


Fig. 2 Using GP models: in addition to the prediction mean value (full line), we obtain a 95% confidence region (gray band) for the underlying function f.

GP models can, like neural networks, be used to model static nonlinearities and can therefore be used for the modelling of dynamic systems [1, 7, 8] as well as time series, if lagged samples of the output signals are fed back and used as regressors. A retrospective review of modelling dynamic systems using GP models can be found in [6].

A dynamic GP model is trained as the nonlinear autoregressive model with an exogenous input (NARX) representation, where the output at time step k depends on the delayed outputs y and the exogenous control inputs u:

$$y(k) = f(y(k-1), \ldots, y(k-n), \\ u(k-1), \ldots, u(k-n)) + \epsilon(k) \tag{10}$$

where f denotes a function, $\epsilon(k)$ is white noise and the output $y(k)$ depends on the state vector $\mathbf{x}(k) = [y(k-1), y(k-2), \ldots, y(k-n), u(k-1), u(k-2), \ldots, u(k-n)]$ at step k.

One of the main limitation of the system identification with GP models is that the computational requirements scales with the third power of the number of training data. In practice this limits the applicability of exact

GP implementation to datasets not exceeding a few thousand data samples. A various computationally efficient approximations to GP models have been proposed as well as on-line modelling [4, 12, 10]. In the next section a streaming-data selection method for GP modelling is proposed that sequentially gains the bulk of the information contained in the streaming data and adjusts with in-coming data. The idea of the selecting data in the proposed method is similar as the idea in Sparse swap algorithm in [5] but used and implemented differently.

3 Streaming-data Selection

In many real-world systems the system's model is, due to its complexity, split into less complex models that are different according to the operating conditions at that moment. Such local models need to be updated on-line, based on incoming information, i.e., streaming data from signal measurements. A possible method for Streaming-Data Selection for GP modelling (SDS-GP) is proposed as follows.

The GP models depend on the training data and the covariance function. In other words, the training data is defined with various regressors and basis functions, while the covariance function is defined with the type and the hyperparameter values. As it is described in Section 2, a GP model's training for a large amount of data is very time consuming. To overcome this greed, only a subset of the most informative data is proposed to be used. Such a subset is, in the literature, called the active set or the basis vectors set [4] and its elements are basis vectors [4], inducing variables [2] or basis functions [9]. With a type or a combination of various types of covariance function, a prior knowledge of the system is included in the model. Nevertheless, with optimisation of the hyperparameter values the model is even better adjusted to the real system. However, in dynamic, non-linear, system identification the squared exponential covariance function is frequently used, presuming the smoothness and stationarity of the system.

The approach we propose processes every new piece of streaming data sequentially, following the Algorithm 1. The main idea of the proposed algorithm is to sequentially gain the bulk of the information contained in the streaming data and to adjust to the dynamic system's characteristic changes. To keep the active set small enough to process the data before new data arrives, the maximum length of the active set should be set with the parameter *max*. Therefore, the parameter *max* is a design parameter. In addition, two more design parameters should be set that determine the "sensitivity" of the algorithm. That means they determine whether new data will be processed or not. The new, incoming data is only processed if the error or the variance of the prediction for new data is higher than the pre-setted thresholds, *thMu* and *thVar*, for the error of the mean value and the variance, respectively.

In other words, the incoming data is only processed if the current model cannot predict the output for the incoming input well enough; otherwise the new data does not contain any new information regarding the current model. Both thresholds can be set heuristically.

Algorithm 1. processData($\mathbf{x}^*$, y^*)

```
data: X,Y,Q
init : max, thMu, thVar
(mu, var) = pred(x*)                                  // get prediction for new data
if abs(mu − y*) > thMu or var > thVar then  // if prediction is not good enough in mean of mean value and variance
    X = X ∪ x*                                  // add new input to the active set
    Y = Y ∪ y*                                  // add new target to the active set
    Q = update(Q, x*)                                   // update inversion
    l = length(X)
    if l > max then                          // if the active set exceeded maximum size
        for i = 1 to l do                    // foreach basis function in the active set
            Q_T = downdate(Q_T, X_j)                    // downdate inversion
            X_T = X \ X_j                    // temporary remove i-th basis input
            Y_T = Y \ Y_j                    // temporary remove i-th basis target
            s(i) = nlml(Q_T, X_T, Y_T)       // calculate neg. log-marginal likelihood
        end
        i_min = argMin(s)                    // get index of the worst basis function
        X = X \ X_imin                       // remove the worst basis input
        Y = Y \ Y_imin                       // remove the worst basis target
        Q = downdate(Q, X_imin)                     // downdate inversion
    end
    hyp = minimise(X, Y, hyp)                // optimise hyperparameter values
    Q = inv(X, Y, hyp)                   // calculate inversion for new hyperparameter values
end
```

In the case of "unknown" data regarding the current model, it is added to the active set. This set actually consists of two sets: the set of inputs $\mathcal{X}$ and the set of targets $\mathcal{Y}$. Furthermore, the inversion of the covariance matrix $\mathbf{Q}$ that is constructed from the $\mathcal{X}$ is extended using a rank-1 update.

If the maximum size of the active set is exceeded, the less informative basis function is removed. The less informative basis function in the active set is found by temporarily eliminating each basis function from the set and calculating the negative log-marginal likelihood for the rest of the basis functions. The rank-1 downdate of the covariance matrix inversion $\mathbf{Q}$ is used, as it is needed to calculate the log-marginal likelihood. The subset with the lowest negative log-likelihood is retained, the corresponding eliminated basis function is removed from the active set and the covariance matrix inversion $\mathbf{Q}$ is appropriately downdated.

For a better adjustment of the model to the real system the hyperparameter values are optimised by maximizing the marginal log-likelihood. This can be done with any suitable optimisation method. In our case it is done with the conjugate gradients optimisation method proposed in [11]. If the hyperparameter values are changed, the covariance matrix has to be updated. In this case the inversion $\mathbf{Q}$ cannot just be updated, but a calculation from scratch is necessary.

It is advisable that the optimisation of the hyperparameter values starts executing after the first few incoming data when the active set contains enough information. Furthermore, the optimisation of the hyperparameter values could be limited to execute only when the difference between the negative log-marginal likelihood of the current and previous step is large enough. In this way some computational load can be preserved.

It should be noted that for the scoring of the incoming data and for the calculating of the hyperparameter values other methods can be used as well.

4 Experiments and Results

To test the viability of the proposed SDS-GP method we performed two experiments where the method is used for sequentially selecting data for the GP modelling from the validation dataset and afterwards used for predicting the validation dataset of the two, non-linear, dynamic systems: *pumadyn-8nm* and *elevators*. Both datasets are publicly available[1] on the web and are often used as benchmarks. The *pumadyn-8nm* dataset is artificially generated using a robot-arm simulator that is highly non-linear and has very low noise. The dataset contains 8192 data samples consisting of 8 regressors each. The first 4096 data samples are used for the training and the other 4096 are used for the validation. The *elevators* dataset relates to controlling the elevators of an F-16 aircraft. It contains two datasets: the first one contains 8752 data samples that are used for the training and 7847 data samples that are used for the validation. Each data sample consists of 18 regressors. It should be noted that all the regressors in both datasets are normalised and that each regressor's values are in the interval $[-1, 1]$.

To test the performance of the SDS-GP method two additional tests were performed on both experiments. In the first test the full GP trained on the entire training dataset is performed. It is used as an indicator of whether or not and how fast the SDS-GP method converges to the full GP model. In the second test the proposed method is compared with an on-line GP modelling method. Among the Sparse On-line Gaussian process (OGP) method [4], the Fast Forward Selection to Speed Up Sparse Gaussian Process Regression

[1] Dataset *pumadyn-8nm* is available at: `http://www.cs.toronto.edu/~delve/data/pumadyn/desc.html` and dataset *elevators* is available at: `http://www.liaad.up.pt/~ltorgo/Regression/elevators.html`

method [12] and the Online Sparse Matrix Gaussian Process Regression and Vision Applications [10] method, the OGP method was chosen as the source code is publicly available[2]. As the OGP method does not optimise the hyperparameter values on-line, it is performed in two modes. In the first mode the $\mathrm{OGP}_{\mathrm{def}}$ default hyperparameter values are initialised as

$$\begin{aligned} w_d &= -2\log(\max(\mathbf{x}_d) - \min(\mathbf{y})/2), \\ v_0 &= \log(\mathrm{var}(\mathbf{y})), \\ v_1 &= \log(\mathrm{var}(\mathbf{y})/4), \end{aligned} \tag{11}$$

where d is the number of regressors. As all the input regressors are normalised, the default values of all w_d equal 0. Therefore, these default hyperparameter values should not be treated as poor values, but rather as neutral values. In the second mode the $\mathrm{OGP}_{\mathrm{opt}}$ optimises the hyperparameter values for the entire training dataset. It should be noted that the SDS-GP hyperparameter values are also initialised using (11) and therefore equal to the $\mathrm{OGP}_{\mathrm{def}}$ hyperparameter values. In this way the importance of the hyperparameter values' optimisation is exposed.

All the tests are compared by the quality of the prediction mean value, which is assessed by computing the mean relative squared error (MRSE)

$$MRSE = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{\sum_{i=1}^{N}\mathbf{y}_i^2}} \tag{12}$$

where $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are the measurement and the prediction in the i-th step. Additionally, the quality of the prediction variance of all the approaches is compared with the logarithm of the predictive density error (LPD)

$$LPD = \frac{1}{2N}\sum_{i=1}^{N}\left(\log(2\pi) + \log(\sigma) + \frac{(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{\sigma_i^2}\right) \tag{13}$$

where σ_i^2 is the prediction variance in the i-th step. While MRSE calculates only the error of the prediction mean value, the LPD also penalises the prediction whose 2^{nd} standard deviation does not cover the real output value.

Discussion

The error measures the MRSE and LPD, depending on the number of basis functions of all the approaches for both experiments, are depicted in Fig. 3 and Fig. 4. It is clear that in both experiments the MRSE of the SDS-GP and $\mathrm{OGP}_{\mathrm{opt}}$ converge to the MRSE of the full GP model quite quickly, while

[2] `http://www.tuebingen.mpg.de/~csatol/ogp/index.html`

the MRSE of the $\mathrm{OGP}_{\mathrm{def}}$ converge more slowly. It should be noted that even though the SDS-GP's initial hyperparameter values are not optimal, with on-line optimisation the MRSE of SDS-GP converge to the MRSE of the full GP.

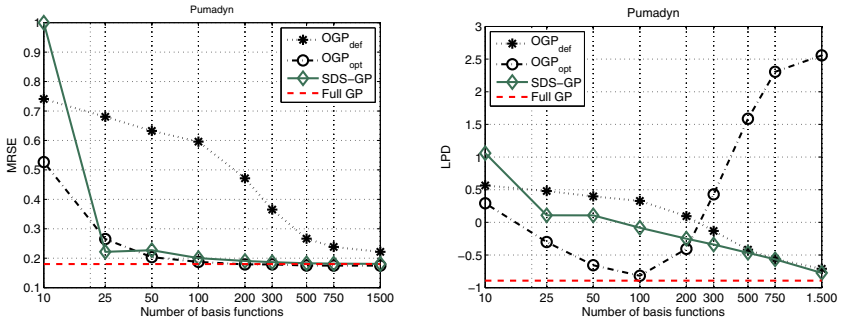


Fig. 3 Case *pumadyn-8nm*: Two error measures depending on the number of basis functions. MRSE is depicted on the left plot and LPD on the right plot.

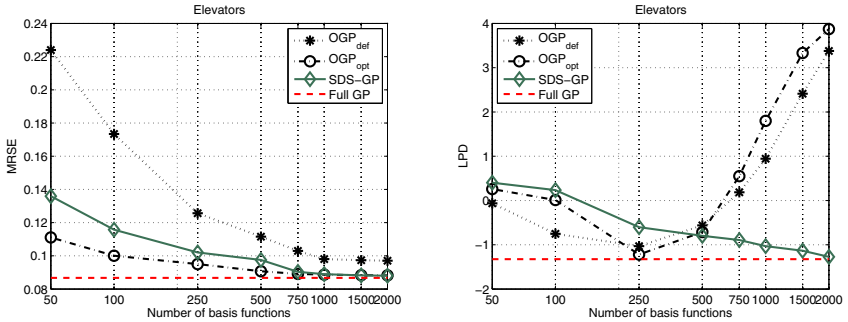


Fig. 4 Case *elevators*: Two error measures depending on the number of basis functions. MRSE is depicted on the left plot and LPD on the right plot.

On the other hand, the LPD of $\mathrm{OGP}_{\mathrm{opt}}$ converge faster than the LPD of SDS-GP and $\mathrm{OGP}_{\mathrm{def}}$, but only for smaller active sets. In the *pumadyn-8nm* case this margin is at about 100 basis functions, while in the *elevators* case it is about 250 basis functions. For larger active sets the LPD of SDS-GP and $\mathrm{OGP}_{\mathrm{def}}$ further converge to the LPD of the full GP, while the LPD of the $\mathrm{OGP}_{\mathrm{opt}}$ starts to diverge. To expose the reason for this occurrence the real output and predictions in the selected region of all the approaches for both experiments are depicted in Fig. 5 and Fig. 6, respectively. For the experiment *pumadyn-8nm* the region from 5457 to 5467 data samples and

models with 750 basis functions are selected and the region from 1925 to 1945 data samples and models with 2000 basis functions for the case *elevators*. It can be seen that the $\mathrm{OGP}_{\mathrm{opt}}$ is over-confident in both cases, while the $\mathrm{OGP}_{\mathrm{def}}$ is only in the *elevators* case. That means it is too self-confident, especially for predictions whose error of the mean value is large and the real output is not inside the 95% confidence region. Such a case can be seen in Fig. 5 for the $\mathrm{OGP}_{\mathrm{opt}}$ and Fig. 6 for both OGP modes, while all the other approaches keep the real output inside the 95% confidence region. Therefore, the LPD of the OGP is higher than the LPD of the other methods.

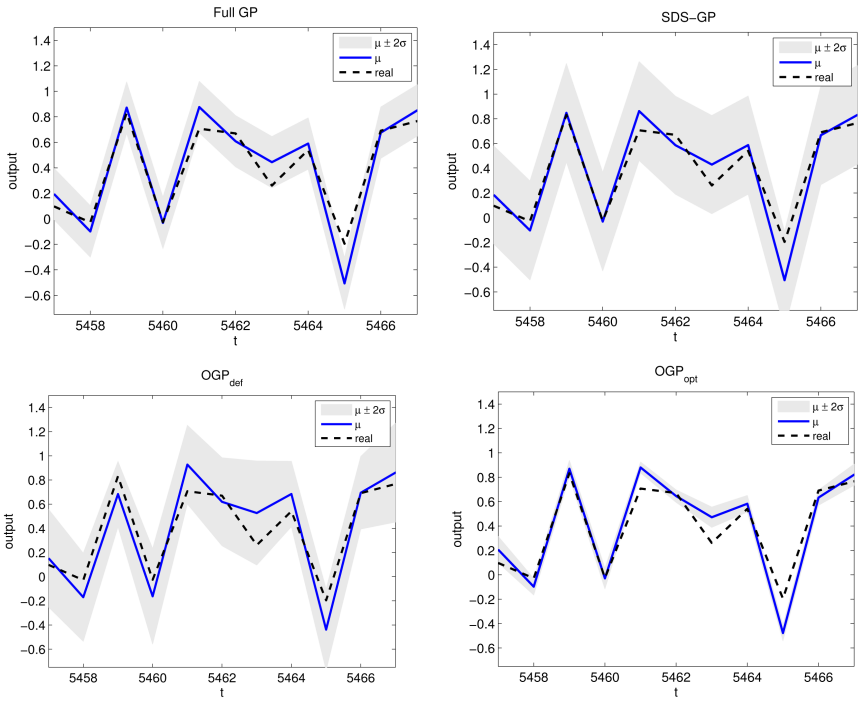


Fig. 5 Case *pumadyn-8nm*: Real output and predictions of all the approaches with 750 basis functions in the region from 5457 to 5467 data samples.

It can be concluded from the presented tests that the SDS-GP method is more suitable for the intended purpose, which is the selection of the incoming data with a rich information content. The selected data is to be used afterwards used for the streaming-data-based modelling of dynamic systems.

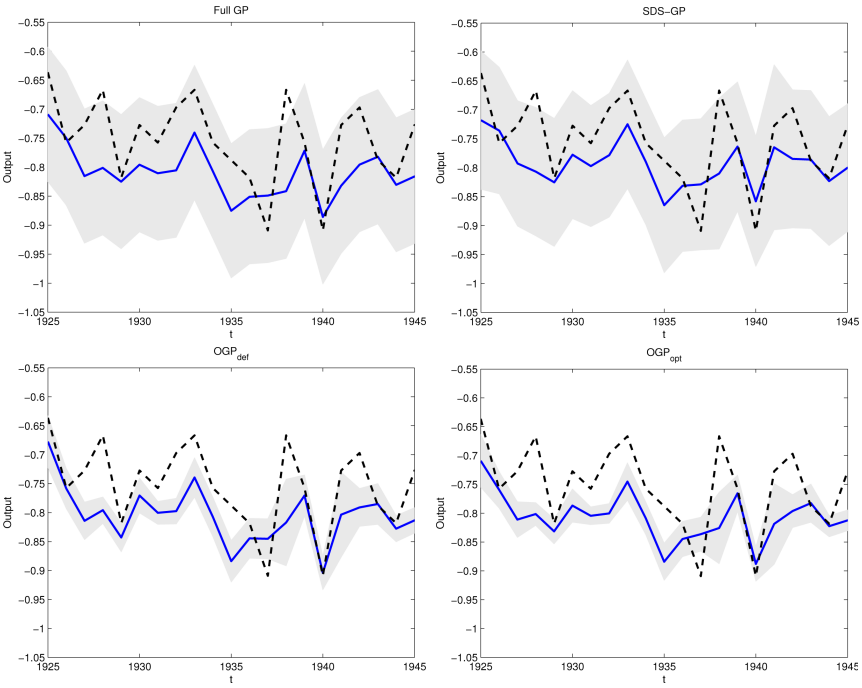


Fig. 6 Case *elevators*: Real output and predictions of all the approaches with 2000 basis functions in the region from 1925 to 1945 data samples.

5 Conclusion

We introduced the Streaming-Data Selection method for GP modelling so that the size of the active set remains constrained. Furthermore, for better model adjustment to the system the hyperparameter values are optimised as well. To test the viability of the proposed method it is tested on two datasets obtained from nonlinear dynamic systems: *pumadyn-8nm* and *elevators*.

The results from the experimental work indicate that the proposed SDS-GP method is viable. That means its error converges towards the error of the full GP, even though the initial hyperparameter values are not optimal. The error of SDS-GP converges, as expected, more slowly, as it contains information only from selected data samples, while the OGP incorporates information from *all* the data samples. On the other hand, the SDS-GP is not over-confident in the case of a larger active set, as is the case for the OGP.

In the performed experiments only the prediction is used for the comparison of the validated models. A more thorough validation of the obtained dynamic systems' models with simulation tests in addition to the prediction tests are envisaged as the future tasks to confirm the usefulness of the proposed SDS-GP method.

Acknowledgements This work was financed by the Slovenian Research Agency, grants Nos. P2-0001 and J2-2099, and (partially) by the European Science Foundation through COST Action IC0702.

References

1. Ažman K, Kocijan J (2007) Application of Gaussian processes for black-box modelling of biosystems. *ISA transactions* 46(4):443–457
2. Quiñonero Candela J, Rasmussen CE (2005) A Unifying View of Sparse Approximate Gaussian Process Regression. *J. Mach. Learn. Res.* 6:1939–1959
3. Quiñonero Candela J, Rasmussen CE, Williams CKI (2007) *Approximation Methods for Gaussian Process Regression.* Tech. rep., Microsoft Research
4. Csató L, Opper M (2002) Sparse on-line Gaussian processes. *Neural Comput.* 14(3):641–668
5. Deisenroth MP (2010) *Efficient Reinforcement Learning using Gaussian Processes.* PhD thesis, Karlsruhe Institute of Technology
6. Kocijan J (2008) Gaussian process models for systems identification. *Proc. 9th Int. PhD Workshop on Systems and Control: young generation viewpoint.* Izola, Slovenia
7. Kocijan J, Girard A, Banko B, Murray-Smith R (2005) Dynamic systems identification with Gaussian processes. *Mathematical and Computer Modelling of Dynamic Systems* 11(4):411–424
8. Kocijan J, Likar B (2008) Gas-liquid separator modelling and simulation with Gaussian-process models. *Simulation Modelling Practice and Theory* 16(8):910–922
9. Lázaro-Gredilla M, Quiñonero Candela J, Rasmussen CE, Figueiras-Vidal AR (2010) Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research* 11:1865–1881
10. Ranganathan A, Yang MH (2008) Online sparse matrix gaussian process regression and vision applications. *Proc. 10th European Conf. on Computer Vision*, I:468–482. Springer-Verlag, Berlin
11. Rassmusen CE, Williams CKI (2006) *Gaussian Processes for Machine Learning.* MIT Press
12. Seeger M, Williams CKI, Lawrence ND (2003) Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. *9th Int. Workshop on Artificial Intelligence and Statistics.* Society for Artificial Intelligence and Statistics

Change Detection Based on the Distribution of p-Values

Katharina Tschumitschew[1] and Frank Klawonn[1,2]

Abstract Non-stationarity is an important aspect of data stream mining. Change detection and on-line adaptation of statistical estimators is required for non-stationary data streams. Statistical hypothesis tests may also be used for change detection. The advantage of using statistical tests compared to heuristic adaptation strategies is that we can distinguish between fluctuations due to the randomness inherent in the underlying distribution while it remains stationary and real changes of the distribution from which we sample. However, the problem of multiple testing should be taken into account when a test is carried out more than once. Even if the underlying distribution does not change over time, any test will erroneously reject the null hypothesis of no change in the long run if we only carry out the test often enough. In this work, we propose methods which account for the multiple testing issue and consequently improve reliability of change detection. A new method based on the information about the distribution of p-values is presented and discussed in this article as well as classical methods such as Bonferroni correction and the Bonferroni-Holm method.

1 Introduction

One of the most important aspects in data stream analysis is that in most applications the underlying data generating process does not remain static, i.e. the underlying probabilistic model cannot be assumed to be stationary. The changes in the data structure may occur over time. Dealing with non-

[1] Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302 Wolfenbüttel, Germany,
{katharina.tschumitschew,f.klawonn}@ostfalia.de
[2] Bioinformatics and Statistics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany, frank.klawonn@helmholtz-hzi.de

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 191–203.
springerlink.com

stationary data requires change detection and on-line adaptation. Different kinds of non-stationarity have been classified in [1]:

- Changes in the data distribution: the change occurs in the data distribution in general. For instance, the mean or the variance of the data distribution may change over time.
- Changes in concept: here concept change refers to changes of a target variable. A target variable is a variable whose values we try to predict based on the model estimated from the data, for instance, for linear regression concept drift refers to the change of the coefficients of the linear model which is used to predict the target variable. Concept change can be further distinguished in the following way:
  - Concept drift: concept drift describes gradual changes of the concept. In statistics, this is usually called structural drift.
  - Concept shift: concept shift refers to an abrupt change which is also referred to as structural break.

In the following, we do not differentiate between concept drift and shift for two reasons. First of all, in both cases the relation between the predictor attributes and the target variable will be changed anyway. Secondly, we can only observe or sample the data at discrete time points, so that it does not matter whether we interpret the changes between two time points as a discontinuous jump in terms of concept shift or as a smooth transition between two time points which we cannot describe or observe in detail, because we have data between two discrete time points.

Real world applications for non-stationary data can be found for instance in stock market or weather prediction, change of protein structures through mutation or the buying behaviour of customers of an on-line store. Since non-stationary data models significantly affect the accuracy of prediction, the fact of concept drift should be taken into account by on-line learning. Hence the effective treatment of non-stationarity is an important problem in machine learning. Therefore change detection and on-line adaptation for data stream mining techniques are required for non-stationary data streams. Various strategies to handle non-stationarity are proposed, see for instance [6] for a detailed survey of change detection methods. Statistical hypothesis tests may also be used for change detection. Since we are working with data streams, it is required that either the calculations for the hypothesis tests can be carried out in an incremental way or time window techniques should be used. Hypothesis tests could be applied to change detection in two different ways (for detailed survey see [12]):

- Change detection through incremental computation of the tests: by this approach the test is computed in an incremental fashion. For instance, the χ^2-test and the t-test (for precise definitions see for example [10]) render themselves easily to incremental computations (on-line adaptation of these tests is described in [12]). A low p-value for the comparison of the

data distributions at different time points – in the case of the χ^2-test – or comparison of the mean values – in the case of the t-test – would indicate a change in the data stream.
- Time window techniques: by this approach the data stream is divided into time windows. A sliding window can be used as well as non-overlapping windows. In order to detect potential changes, we need either to compare data from an earlier window with data from newer one or to test only the new data (for instance, whether the data follow a known or assumed distribution).

However, the problem of multiple testing should be taken into account when more than one hypothesis is tested simultaneously. The more hypotheses are tested, the more likely the null hypothesis of no change will be erroneously rejected, even if the underlying distribution does not change over time. In this work we present different approaches to solve this problem. One way is the application methods that account for multiple testing like the well known Bonferroni correction and the Bonferroni-Holm method. Furthermore, we propose a new approach based on the information about the distribution of p-values.

This paper is organised as follows. The problem of multiple testing is explained in Section 2. Two classical methods to handle the problem of multiple testing are also described in this section. In Section 3 the theoretical background on p-values is given and a new approach based on the distribution of p-values under the null hypothesis is introduced. Examples are discussed in the experimental section 4.

2 Multiple Testing

Multiple testing refers to the application of a number of tests simultaneously. Instead of a single null hypothesis, tests for a set of null hypotheses H_0, $H_1, \ldots, H_n$ are considered. These null hypotheses do not have to exclude each other.

An example for multiple testing is a test whether m random variables $X_1, \ldots X_m$ are pairwise independent. This means the null hypotheses are $H_{1,2}, \ldots, H_{1,m}, \ldots, H_{m-1,m}$ where $H_{i,j}$ states that X_i and X_j are independent. Multiple testing leads to the undesired effect of cumulating the α-error.

Definition 1. The α-error α is the probability to reject the null hypothesis erroneously, given it is true.

Choosing $\alpha = 0.05$ means that in 5% of the cases the null hypothesis would be rejected, although it is true. When k tests are applied to the same sample, then the error probability for each test is α. Under the assumption that the null hypotheses are all true and the tests are independent, the probability that at least one test will reject its null hypothesis erroneously is

$$\begin{aligned} P(\ell \geq 1) &= 1 - P(\ell = 0) \\ &= 1 - (1-\alpha) \cdot (1-\alpha) \cdot \ldots \cdot (1-\alpha) \\ &= 1 - (1-\alpha)^k. \end{aligned} \tag{1}$$

ℓ is the number of tests rejecting the null hypothesis.

A variety of approaches have been proposed to handle the problem of cumulating the α-error. In the following, two common methods will be introduced shortly.

The simplest and most conservative method is Bonferroni correction [9]. When k null hypotheses are tested simultaneously and α is the desired overall α-error for all tests together, then the corrected α-error for each single test should be chosen as $\tilde{\alpha} = \frac{\alpha}{k}$. The justification for this correction is the inequality

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i). \tag{2}$$

For Bonferroni correction, A_i is the event that the null hypothesis H_i is rejected, although it is true. In this way, the probability that one or more of the tests rejects its corresponding null hypothesis is at most α. In order to guarantee the significance level α, each single test must be carried out with the corrected level $\tilde{\alpha}$.

Bonferroni correction is a very rough and conservative approximation for the true α-error. One of its disadvantages is that the corrected significance level $\tilde{\alpha}$ becomes very low, so that it becomes almost impossible to reject any of the null hypotheses.

The simple single step Bonferroni correction has been improved by Holm [7]. The Bonferroni-Holm method is a multi-step procedure in which the necessary corrections are carried out stepwise. This method usually yields larger corrected α-values than the simple Bonferroni correction.

When k hypotheses are tested simultaneously and the overall α-error for all tests is α, for each of the tests the corresponding p-value is computed based on the sample x and the p-values are sorted in ascending order.

$$p_{[1]}(x) \leq p_{[2]}(x) \leq \ldots \leq p_{[k]}(x) \tag{3}$$

The null hypotheses H_i are ordered in the same way.

$$H_{[1]}, H_{[2]}, \ldots, H_{[k]} \tag{4}$$

In the first step $H_{[1]}$ is tested by comparing $p_{[1]}$ with $\frac{\alpha}{k}$. If $p_{[1]} > \frac{\alpha}{k}$ holds, then $H_{[1]}$ and the other null hypotheses $H_{[2]}, \ldots, H_{[k]}$ are not rejected. The method terminates in this case. However, if $p_{[1]} \leq \frac{\alpha}{k}$ holds, $H_{[1]}$ is rejected and the next null hypothesis $H_{[2]}$ is tested by comparing the p-value $p_{[2]}$ and the corrected α-value $\frac{\alpha}{k-1}$. If $p_{[2]} > \frac{\alpha}{k-1}$ holds, $H_{[2]}$ and the remaining null

hypotheses $H_{[3]}, \ldots, H_{[k]}$ are not rejected. If $p_{[2]} \leq \frac{\alpha}{k-1}$ holds, $H_{[2]}$ is rejected and the procedure continues with $H_{[3]}$ in the same way.

The Bonferroni-Holm method tests the hypotheses in the order of their p-values, starting with $H_{[1]}$. The corrected α_i-values $\frac{\alpha}{k}, \frac{\alpha}{k-1}, \ldots \alpha$ are increasing. Therefore, the Bonferroni-Holm method rejects at least those hypotheses that are also rejected by simple Bonferroni correction, but in general more hypotheses can be rejected.

During change detection instead of the common significance level α, the Bonferroni correction or Bonferroni-Holm method should be used in order to avoid the multiple testing problem. However, the streaming nature of the data should be taken into account and it is therefore impossible to hold all the obtained p-values in the memory. Furthermore, the number of tests to be carried out is not known in advance. Thus, a time window technique-based approach should be used, such for instance as a sliding window or non-overlapping time windows.

3 Meta p-values

Another possibility to solve the problem of multiple testing during change detection is to study the behaviour of the obtained p-values. Several authors have analysed properties of p-values. For instance, Gibson and Pratt (see [5]) provided an interpretation and methodology for p-values, Sackrowitz and Samuel-Cahn [8] analysed the stochastic behaviour of p-values. Donahue in [4] studied the distribution of p-values under the alternative hypothesis. In [2], the authors focus on the median of the p-value under the alternative hypothesis.

Definition 2. The p-value is the probability to obtain a value of the test statistic as extreme as, or more extreme than (depending on the alternative hypothesis) the observed value of the test statistic given the null hypothesis is true.

Hence, in the case of continuous test statistics for a right tailed test the p-value is calculated as

$$p = \Pr\left(T \geq t | H_0\right) = 1 - F_T(t) \tag{5}$$

and for a left tailed test as

$$p = \Pr\left(T \leq t | H_0\right) = F_T(t) \tag{6}$$

where $F_T(t)$ is the cumulative distribution function for the test statistic T under the assumption that the null hypothesis H_0 is true.

In the case of a two tailed test, the p-value is the total area under both tails with an area of $\frac{p}{2}$ in each tail. Therefore, if the observed value falls into

the one-tailed area, the area of this tail has to be doubled and the other tail can be ignored.

$$p = \begin{cases} 2 \cdot \Pr\left(T \leq t | H_0\right), \text{ if } t \leq q^T_{0.5} \\ 2 \cdot \Pr\left(T \geq t | H_0\right), \text{ otherwise.} \end{cases} \tag{7}$$

As the Equations (5), (6) and (7) show, the p-value is a function of a random variable and hence a random variable itself. An obvious question is: how are the p-values distributed under the null hypothesis and how under the alternative hypothesis? First, the distribution of p-values is analysed when H_0 is true (see [4, 8]).

Theorem 1. *Given the null hypothesis is true, the p-values of a continuous test statistic T follow a uniform distribution on the unit interval $[0, 1]$.*

Proof. Let p be the achieved p-value and t the calculated test statistic with $F_P(p|H_0)$ and $F_T(t)$ being the corresponding cumulative distribution functions under H_0. Also, let $F_T^{-1}(\gamma)$ be the inverse function of $F_T(t)$, so that $F_T\left(F_T^{-1}(\gamma)\right) = \gamma$ for all $\gamma \in [0, 1]$. Then, for a right tailed test the following holds

$$\begin{aligned} F_P(p|H_0) &= \Pr\left(P \leq p | H_0\right) \\ &= \Pr\left(1 - F_T\left(t\right) \leq p | H_0\right) \\ &= \Pr\left(F_T\left(t\right) \geq (1-p) \,|H_0\right) \\ &= 1 - \Pr\left(F_T\left(t\right) \leq (1-p) \,|H_0\right) \\ &= 1 - F_T\left(F_T^{-1}\left(1-p\right)\right) \\ &= 1 - (1-p) = p \end{aligned} \tag{8}$$

For a left tailed test corresponding to Equation (6) the distribution function of the p-value is as follows

$$\begin{aligned} F_P(p|H_0) &= \Pr\left(P \leq p | H_0\right) \\ &= \Pr\left(F_T\left(t\right) \leq p | H_0\right) \\ &= F_T\left(F_T^{-1}\left(p\right)\right) \\ &= p \end{aligned} \tag{9}$$

for all $p \in [0, 1]$.

According to Equations (7), (8) and (9), we obtain for the distribution of P in case of a two tailed test: $F_P(p|H_0) = 2 \cdot \frac{p}{2} = p$. Note that we divide the probability p to equal parts between both tails. Therefore, the random variable P is uniformly distributed on the interval $[0, 1]$ when H_0 is true. □

Figures 1 and 2 show the histograms for simulated p-values under the null hypothesis and the alternative hypothesis respectively. The p-values are generated by the Kolmogorov-Smirnov test which is carried out over and over again for the problem of testing test whether or not data are coming from a standard normal distribution.

In the case when alternative hypothesis is true, the data are generated by a normal distribution with expected value 0.05 and standard deviation 1. Altogether, 100 different runs are made for data samples of length 1000. Figure 1 confirms that the p-values follow a uniform distribution on the unit interval $[0, 1]$ when the null hypothesis is true.

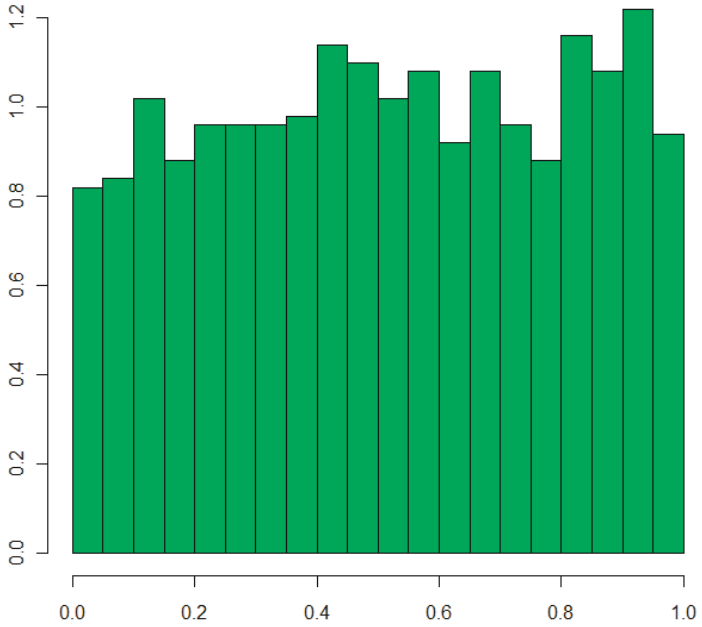


Fig. 1 Histogram for p-values under the null hypothesis.

The histogram in Figure 2 for the alternative hypothesis shows a different situation. The sampling distribution here is clearly not uniform anymore, the majority of values is close to zero and the amount of values decreases towards the p-value one.

Thus, we are interested in the question: how are p-values distributed when the alternative hypothesis holds? The distribution is given by Equation (10) (see [4]).

$$\begin{aligned} F_P(p|H_1) &= \Pr\left(P \leq p|H_1\right) \\ &= \Pr\left(1 - F_T\left(t\right) \leq p|H_1\right) \\ &= \Pr\left(F_T\left(t\right) \geq (1-p)\,|H_1\right) \\ &= 1 - \Pr\left(F_T\left(t\right) \leq (1-p)\,|H_1\right) \\ &= 1 - G_T\left(F_T^{-1}\left(1-p\right)\right) \end{aligned} \tag{10}$$

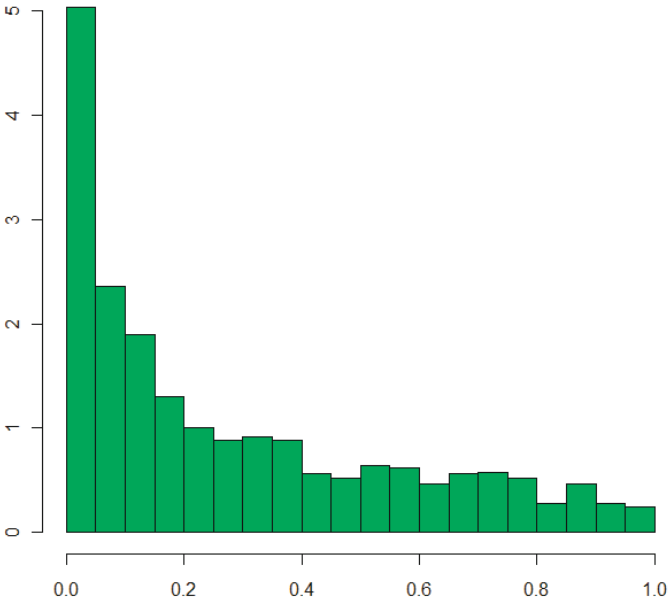


Fig. 2 Histogram for p-values under the alternative hypothesis.

where G_T is the distribution of the test statistic T under the alternative hypothesis. Here we only consider upper-tailed one-sided tests. As Equation (10) shows, the distribution of the p-values in this case depends on the test statistic distribution under H_0 as well as under H_1 hypothesis.

Hence, knowing the distribution of p-values under both hypotheses, a meta analysis can be performed. Since for each alternative hypothesis – in most cases the alternative is a composite hypothesis representing not a single but a set of distributions – and therefore for each G_T the distribution of p-values under H_1 is different (see Equation (10)) we restrict further considerations to the uniformity of p-values under H_0.

The most obvious way to carry out a meta analysis is to perform a goodness of fit test on the obtained p-values during multiple testing. For instance, the Kolmogorov-Smirnov test (an implementation is available in the R statistics library [3]) can be used for that purpose. However, the following problem should be taken into account: in order to carry out a meta analysis of p-values, neither a sliding window nor an incremental computation can be used for change detection. Indeed, the general assumption for hypothesis tests that the considered random variables are independent and identically distributed (i.i.d.) does not hold for overlapping sliding windows. By the application of sliding windows or incremental computation the next p-value is highly dependent on the previous ones. The reason for this problem is that almost the same values are used by the hypothesis test, correspondingly the com-

puted neighbouring p-values would be approximately equal. Therefore, for this approach only non-overlapping windows should be used during change detection. As a consequence, we can not use the comparison between data from an earlier window with data from newer one when an abrupt change occurs, since in such a case H_0 would be false only once and therefore only one p-value would not come from a uniform distribution. Nevertheless, this approach shows good results when a test is used in order to proof whether the data follow a known or assumed distribution or to detect drift in the data generating process.

4 Experimental Results

Our approach has been implemented in Java using R-libraries and has been tested with artificial data. For the data generation process the following model was used: first n_1 time points data are generated from a standard normal distribution, i.e. $X_i \sim N(0,\ 1)$ for $i \in \{1, \ldots, n_1\}$. At time point $n_1 + 1$ a change occurs and the data are normally distributed with the following settings: $\mu = 0.1$ and $\sigma = 1$, i.e. $N(0.1,\ 1)$.

Our meta analysis of p-values has been applied to this data set. The Kolmogorov-Smirnov test for standard normality of the data was carried out for non-overlapping time windows. The size of the window for the change detection was chosen to be 500. Afterwards, the sliding window of size 100 was used for the meta analysis of the obtained p-values. In order to test the distribution of the p-values a Kolmogorov-Smirnov test for uniformity is used. A meta p-value is consequently the result of this test. Figure 3 illustrates described technique.

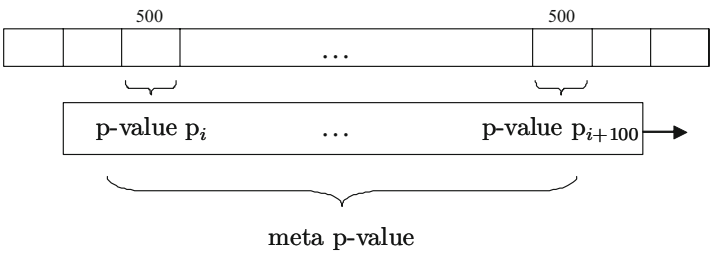


Fig. 3 Two windows for change detection.

The change occurred at the time point 59489. The computed p-values for this part of the data are as follows:

After the change occurs, the null hypothesis can be rejected (depending on the chosen α). However, from time to time H_0 cannot be rejected. Furthermore, for some parts without change H_0 is erroneously rejected, even

Table 1 p-values obtained during change detection.

time window	p-value
[57000; 57499]	0.07486618716777954
[57500; 57999]	0.1401818038913014
[58000; 58499]	0.9941898244705528
[58500; 58999]	0.9291249862216258
[59000; 59499]	0.5020298421810007
[59500; 59999]	0.01168733233091191
[60000; 60499]	0.05625967117647695
[60500; 60999]	0.6789664978854166
[61000; 61499]	0.394486208210243
[61500; 61999]	0.05360718854238174
[62000; 62499]	0.7747463214977733

though the underlying distribution did not change at that time. For instance for the interval [45500; 45999] the p-value is 0.018673 and consequently H_0 can be rejected for all $\alpha \geq 0.020$. Whereas as Table 2 shows, all meta p-values are smaller than 0.05 starting from the window [41000; 65999] and all meta p-values before are larger than 0.05.

Table 2 Meta p-values obtained during change detection.

time window	meta p-value
[40000; 64999]	0.1599219
[40500; 65499]	0.0809654
[41000; 65999]	0.0377086
[41500; 66499]	0.0161466
[42000; 66999]	0.0063506
[42500; 67499]	0.0022917

For the next example the data were generated as follows:

$$Y_t = \sum_{i=1}^{t} |X_i| \,. \tag{11}$$

We assume the random variables X_i to be normally distributed with expected value $\mu = 0$ and variance σ_1^2, i.e. $X_i \sim N\left(0,\ \sigma_1^2\right)$. To make the situation more realistic, we consider the following model:

$$Z_t \sim N\left(y_t, \sigma_2^2\right). \tag{12}$$

The process (12) can be understood as a constant model with drift and noise. The noise follows a normal distribution whose expected value equals the actual value of the random walk and whose variance is σ_2^2. The data were generated with the following parameters: $\sigma_1 = 0.00000008, \sigma_2 = 0.002$. Figure 4 shows the generated data.

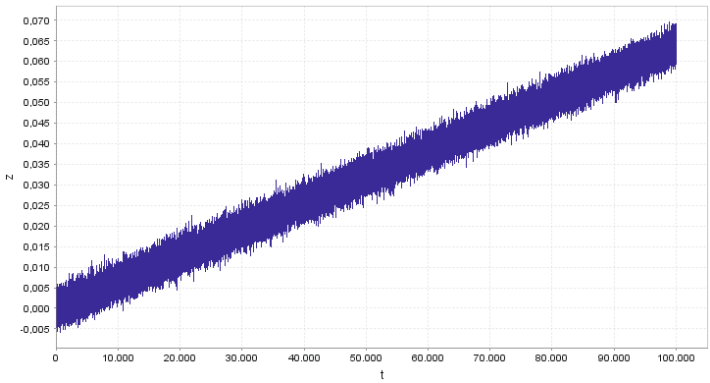


Fig. 4 Two windows for change detection.

In order to detect changes, the two sample t-test was applied to this data set. In such a way we can test whether the data from the old and new windows have the same mean. Two non-overlapping windows of size 500 are used. For the meta analysis, similar as before, the Kolmogorov-Smirnov test for uniformity is applied to a sliding window of size 50. Since the mean changes very slightly, sometimes H_0 can not be rejected (depending on the chosen α), as can be seen from Table 3, whereas all meta p-values provide the strong evidence that the data is non-stationary (all obtained meta p-values are smaller than 10^{-9}).

As Tables 1, 2 and 3 show, the meta p-values are more reliable than p-values. However, it should be taken into account that more time is needed until a change can be detected. Therefore, this approach is not suitable when very fast reaction to the occurred change is required. Whereas when more attention is paid to the accuracy of change detection, meta p-values provide a good solution to the problem of multiple testing for non-stationarity of the data. For instance such kind of change detection can be used for changes caused by slow wear and abrasion of materials, here the fast reaction is not required but the information about the speed of wear.

Table 3 p-values obtained during change detection.

time window t	time window $t+1$	p-value
[0; 499]	[500; 999]	0.04019542802527917
[500; 999]	[1000; 1499]	0.01835982391226245
[1000; 1499]	[1500; 1999]	0.02694198995888841
[1500; 1999]	[2000; 2499]	0.00590771301502357
[2000; 2499]	[2500; 2999]	0.00000051742252253
[2500; 2999]	[3000; 3499]	0.21019670543166669
[3000; 3499]	[3500; 3999]	0.02610004716763162
[3500; 3999]	[4000; 4499]	0.01388767893804595
[4000; 4499]	[4500; 4999]	0.02986063639554551
[4500; 4999]	[5000; 5499]	0.00174724618341983
[5000; 5499]	[5500; 5999]	0.21651140022620408
[5500; 5999]	[6000; 6499]	0.00180512145155431

5 Conclusion

Change detection is a crucial aspect for non-stationary data streams or "evolving systems". It has been demonstrated in [11] that naïve adaption without taking any effort to distinguish between noise and true changes of the underlying sample distribution can lead to very undesired results. Statistical measures and tests can help to discover true changes in the distribution and to distinguish them from random noise. However, the following problem arises: when a test is carried out over and over again, the probability to erroneously rejecting the null hypothesis increases with the amount of applied tests. In this work, we have discussed the problem of multiple testing during change detection and proposed classical methods as well as a new approach to cope with the multiple testing issue.

Bonferroni correction and the Bonferroni-Holm method adjust the significance level α in order to correct the occurrence of incorrect rejections of H_0 leading to a very conservative approach that will seldom indicate a change in the data stream. Our proposed approach is based on the uniformity of the p-values under the null hypothesis. In such a way, not only the p-values but also the meta p-values are taken into account by the change detection. This approach shows good results even in cases where Bonferroni correction and the Bonferroni-Holm method could not achieve any improvement. Although we have only considered the distribution of the p-values under the null hypothesis, it could be useful to study the distribution of p-values under the alternative hypothesis, too.

References

1. Basseville M, Nikiforov I (1993) *Detection of Abrupt Changes: Theory and Application*. Prentice Hall Prentice Hall, Upper Saddle River, New Jersey (1993)
2. Bhattacharya B, Habtzghi D (2002) Median of the p value under the alternative hypothesis. *The American Statistician* 56:202–206
3. Crawley M (2005) *Statistics: An Introduction using R*. J. Wiley & Sons, New York
4. Donahue RMJ (1999) A note on information seldom reported via the P value. *The American Statistician* 53(4):303–306
5. Gibbons J, Pratt J (1975) p-values: Interpretation and methodology. *The American Statistician* 29:20–25
6. Gustafsson F (2000) Adaptive Filtering and Change Detection. J. Wiley & SOns, New York
7. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65–70
8. Sackrowitz H, Samuel-Cahn E (1999) p-values as random variables—expected p-values. *The American Statistician* 53(4):326–331
9. Shaffer JP (1995) Multiple hypothesis testing. *Ann. Rev. Psych* 46:561–584
10. Sheskin D (1997) *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC-Press, Boca Raton
11. Tschumitschew K, Klawonn F (2010) The need for benchmarks with data from stochastic processes and meta-models in evolving systems. *Int. Symp. Evolving Intelligent Systems*, 30–33. SSAISB, Leicester
12. Tschumitschew K, Klawonn F (2012) Incremental statistical measures. In: Sayed-Mouchaweh M, Lughofer E (eds.) *Learning in non-stationary environments: Methods and Applications*, Chap. 2. Springer, New York

Advanced Analysis of Dynamic Graphs in Social and Neural Networks

Pascal Held[1], Christian Moewes[1], Christian Braune[1], Rudolf Kruse[1], and Bernhard A. Sabel[2]

Abstract Dynamic graphs are ubiquitous in real world applications. They can be found, e.g. in biology, neuroscience, computer science, medicine, social networks, the World Wide Web. There is a great necessity and interest in analyzing these dynamic graphs efficiently. Typically, analysis methods from classical data mining and network theory have been studied separately in different fields of research. Dealing with complex networks in real world applications, there is a need to perform interdisciplinary research by combining techniques of different fields. In this paper, we analyze dynamic graphs from two different applications, i.e. social science and neuroscience. We exploit the edge weights in both types of networks to answer distinct questions in the respective fields of science. First, for the representation of edge weights in a social network graph we propose a method to efficiently represent the strength of a relation between two entities based on events involving both entities. Second, we correlate graph measures of electroencephalographic activity networks with clinical variables to find good predictors for patients with visual field damages.

1 Introduction

Complex dynamic networks are ubiquitous. They can be found, e.g. in biology [11], neuroscience [28], computer science [10], medicine [24], social networks [16], and the World Wide Web [14]. There is a great necessity and interest in analyzing these dynamic graphs efficiently as patterns inside of

[1] Working Group on Computational Intelligence, Faculty of Computer Science, Otto-von-Guericke University, Magdeburg, Germany, {pheld,cmoewes,kruse}@ovgu.de, christian.braune@st.ovgu.de

[2] Institute of Medical Psychology, Medical Faculty, Otto-von-Guericke University, Magdeburg, Germany, bernhard.sabel@med.ovgu.de

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 205–222.
springerlink.com

these structures might reveal knowledge about the underlying system. Classically, analysis methods from both network theory and knowledge discovery in databases have been studied separately in different fields of research. The analysis of complex networks as they occur in real world applications can be supported by combining techniques of these two fields [38, 17]. In this paper, we present two real-world problems of dynamic graphs from distinct applications, i.e. social science and neuroscience. We exploit the edge weights in both types of networks to answer distinct questions in the respective fields of science. For the social science problem we propose a method to efficiently represent the strength of a relation between two entities based on events involving both entities. For the neuroscience problem we show how electroencephalographic (EEG) activity networks of patients suffering from visual field defects can be correlated with clinical variables for feature selection. Parts of the latter study have already been orally presented at a workshop [19].

2 Social Network Analysis

Social network analysis has already been popular long before websites like Facebook, XING or Google+ — now commonly understood/known as social networks — were launched. In [33] a comprehensive approach of modeling social network data as (un)directed graphs has been proposed and has been widely accepted. Over the years a lot of research has been performed on e.g. cohesiveness of groups of members in social graphs [36] or segmentation of social networks [16]. All these methods have in common that they use a static representation of the social graph underlying the respective social network.

Attempts have been made to infer information from dynamic graphs (e.g. in [1]) but they either restrict themselves to fairly simple questions like connectivity or to path finding problems in order to cope with the changing structure of the graph. These approaches suffer from being discretized images of an originally continuously changing structure. Such discretization results from some kind of binning operation performed on the data, thus leading to a loss of information, namely the exact time when an event has happened. Such an approach does not take into account the frequency with which events occur but rather lists their absolute number.

The Butterworth filter [3] is one of the best-known infinite impulse response filters. One of its most interesting features is its flat frequency response, i.e. it does not generate rippling effects, when the signal strength changes. Interpreting the binned events of a social graph as a time- and strength-discretized signal the filter response of such a Butterworth filter should have the desired properties that events (dirac pulses) can be binned while keeping some information on the frequency.

2.1 Butterworth Filtering

Representing the structure of a social network not only by the *friendship* relation (i.e. nodes represent persons, edge if they befriended each other), which results in a more or less static description of the graph, but also by adding weights to such edges where the weight reflects the amount of activity between the two corresponding nodes, requires a way to describe this activity. Event-based weighting of edges in a social graph could be accomplished by simply storing all the timestamps at which events between two nodes occurred. Obviously this approach would become unfeasible very soon due to the amount of memory required for such a procedure. An additional disadvantage of such an approach would be that, while we can make statements about the point in time when an event occurred. If possible at all, we can roughly estimate the current weight that should be assigned to an edge at a given point in time. Operations like a sliding average would be able to adapt to such a problem with the major drawback, that only a small time frame can be used to determine the current average due to memory restrictions — no further information about the past is available if only such a value is used.

From electronic signal processing the Butterworth filter is a well-known variant of an infinite impulse response filter that produces an output signal as response to its input signal without causing the rippling effects from which other filters suffer. In general such a filter is defined by two sets of coefficients B and A and the filter's response y for a signal x at the bin n can be obtained by computing

$$y_n = \sum_{i=1}^{n_b} (b_i \cdot x_{n-(i-1)}) - \sum_{j=2}^{n_a} (a_j \cdot y_{n-(j-1)}), \text{ where}$$

$$\{b_1, \ldots, b_{n_b}\} = B \text{ and } \{a_1 = 1, a_2, \ldots, a_{n_a}\} = A.$$

This recursive representation makes it possible to avoid enumerating all signal values from negative to positive infinity.

Other parameters that either influence the shape of the resulting curve or the set of possible edges that are considered are:

- Step width: Amount of milliseconds falling into one time bin
- Grade: The grade of the filter determines the two sets B, A of coefficients responsible for the shape of the resulting filter response. The number of coefficients depends directly on the grade and describes how many past signal (and response) values are considered for the calculation.
- Minimum messages: During preprocessing of the data only certain edges were included in the graph depending on how many messages were sent in total.

2.2 Data Sets

For the analysis and validation of our method we used the well-known Enron data set[1]. We removed both external contacts from the data (Enron employees sending mails to non-Enron employees) and all mail contacts with mailing lists. Duplicates (*firstname.lastname* vs. *firstnamelastname*) have been reduced to one single node and mails that were sent to several users at once were treated as separate events (such that a mail sent from A to B and C was considered as two identical mails that were sent from A to B and from A to C).

2.3 Applying the Filter to the Data

As the Butterworth filter produces a continuous signal we want the filter response to be $1/l$ for a time step of length l to give a better generalization. This restriction and the fact that it produces an equal sum of values over a continuous time span directly lead to two adversing goals in finding an optimal frequency to describe the filter:

1. Minimize the difference between the discretely binned signal and the filter response.
2. Find a frequency $f \in (0, 1)$ that produces a continuous, smooth and nearly linear approximation of the signal (i.e. has only a few local extrema).

While the number of extrema can be reduced by lowering the passband frequency (which at some point will result in a nearly constant response), the error can be reduced by increasing it. This interrelation is illustrated by Fig. 1, which shows the filter response for three different passband frequencies when applied to event data from the Enron data set (for simplicity, the data were treated as if belonging to a single edge). Of course the data should be split up into the real edges for any further analysis.

To evaluate the total complexity of the resulting model depending on the frequency we adapted Akaike and Bayesian Information Criterion (AIC / BIC) [2, p. 110] to include the parameters we want to optimize on. For any given frequency f we can compute the mean squared error (MSE) for the resulting signal and count the number of extrema. The number n_e of extrema can be used as a measure for the complexity of the resulting curve by assuming we have to store this curve as a polynomial with a degree of $n_e + 1$.

Thus, the objective functions we need to minimize are

$$AIC(f) = 2k - 2 \cdot \ln(L) \quad \text{and} \quad BIC(f) = k \cdot \ln(n_e) - 2 \cdot \ln(L),$$

[1] Obtained from http://www.cs.cmu.edu/~enron/.

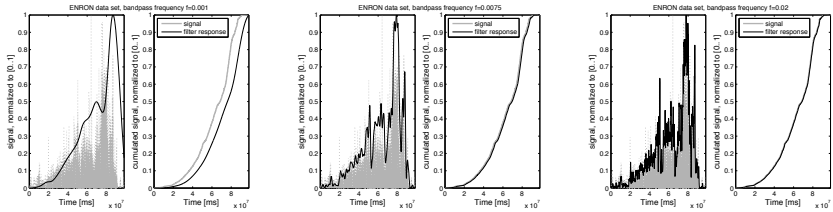


Fig. 1 Filter response for different passband frequencies for Enron data set, time binning: 10'000 $^{\text{ms}}/_{\text{bin}}$.

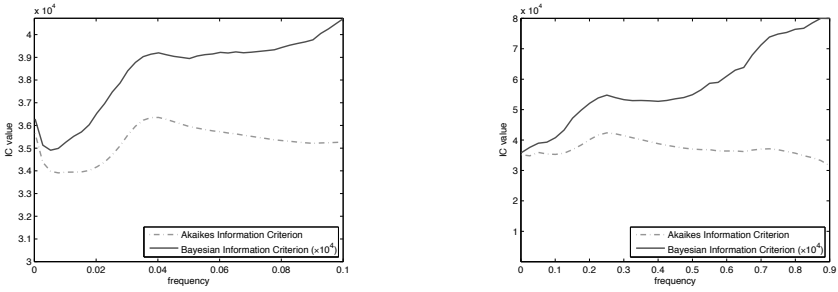


Fig. 2 Left: AIC and BIC for the Enron data set, plotted against different passband frequencies (filter grade = 4). Right: Development of the two information criteria used over a larger interval. The interval $(0.9, 1]$ has been left out as the MSE heads towards zero in these cases which in turn leads to a term in AIC becoming negative.

respectively, where k is the number of parameters and L is the likelihood of the model. Assuming the error in the model is normally distributed both functions can be simplified to

$$AIC(f) = 2k + n \cdot \ln(MSE) \quad \text{and} \quad BIC(f) = k \cdot \ln(n) + n \cdot (MSE).$$

The MSE for a frequency f can be computed from the signal X and the filter response y as

$$MSE(f, X) = \frac{1}{\|X\|} \sum_{i=1}^{\|X\|} (x_i - y_i)^2,$$

the number of parameters equals $n_e + 2$.

These functions can then be optimized using standard optimization techniques like simulated annealing [13] or gradient descent [27] techniques. According to the resulting shape of the curves for both objective functions (see Fig. 2) the optimal passband frequency for the depicted example data set lies near $f = 0.0075$, when limited to $(0, 0.2]$. Higher frequencies result in a filter response that does never fulfill the smoothness requirement although they might lead to lower values of the objective functions (see also Sect. 2.4).

For the purpose of storing event information in a coherent way across multiple edges in the interaction graph it is useful to only use one global frequency to apply the same filter on all edges. This removes the need to store the individual filter parameters, and results in only storing the last few signal and filter response values to be able to calculate the new filter response with the given, global parameter set.

Such a multi signal optimization can be performed if not only the MSE of an individual signal is calculated for each frequency but the MSE over all edges of the graph. The number of local extrema k_{avg} is equally easy to obtain as the average number of extrema contained over all edges. With this the multi signal objective functions are

$$AIC(f) = 2k_{avg} + n \cdot \ln(MSE_{global}) \quad \text{and}$$
$$BIC(f) = k_{avg} \cdot \ln(n) + n \cdot (MSE_{global}),$$

respectively, where the

$$MSE_{global}(f, X) = \frac{1}{\sum_{x \in X} \|x\|} \cdot \sum_{x \in X} \sum_{i=1}^{\|x\|} (x_i - y(i))^2$$

can be considered a MSE over all considered bins.

2.4 Evaluation

Naturally, when compared to a moving average filter, the MSE of our approach as compared to the original signal will be significantly higher (see Tbl. 1). As we never aimed at solely minimizing the error but also the complexity of the response signal, our method outperforms the moving average when using the BIC as optimization criterion. Though it may seem that the moving average performs better when considering AIC this is owed to this measure being biased toward models with very high complexity. This effect can be seen on the right-hand side in Fig. 2. Here can be seen clearly, that the AIC has its true minimum for a frequency above 0.8. Hence we restricted optimization already to find an optimum only in the interval $(0, 0.2]$ where desirable (in terms of local extrema) results are achieved.

	Moving Average	Butterworth Filter
MSE	19.3163	33.3691
AIC	31068	33445
BIC	3.8814	3.4367

Table 1 Different evaluation measures to compare the moving average with the Butterworth Filter for the Enron data set.

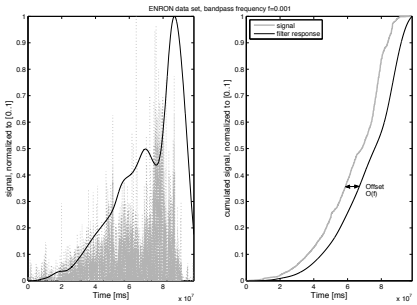


Fig. 3 Offset between the true (left) and the response signal in the cumulated signal function (right).

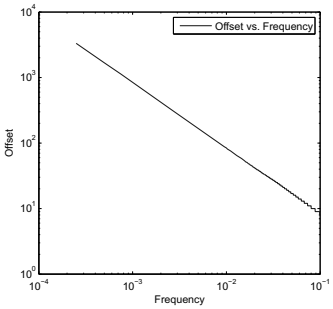


Fig. 4 Optimal offset vs. passband frequency of Butterworth filter obtained by all experiments.

As already described above, we observed during the evaluation of the error measures, that depending on the passband frequency the filter response shows some offset (see Fig. 2.4a), which decreases with increasing frequency. We tried to find a best offset which could be applied to the filter response in order to reduce the overall error that occurs simply due to the offset. Plotting these offsets against the frequency they correspond to leads to Fig. 2.4b.

Simple curve fitting yields that the optimal offset $o(f)$ can be calculated directly from the passband frequency used by the filter with the following formula: $o(f) = \left\lceil \frac{a}{f} + b \right\rceil$, where $a = 0.8347$ $[0.8341, 0.8352]$ and $b = 0.3388$ $[0.2085, 0.4691]$, (in brackets 95% confidence bounds). Actually the exponent for the factor f is not -1 but it is so close that we fixed it at -1 for simplicity. As we only have discrete bins, such a simplification seems reasonable as the following discretization of the result will obliterate most imprecisions. All of our experiments show that this formula seems to be independent from the given data set. That led us to the assumption to introduce this as a correction term into the objective function. This may be an important step for scenarios where the behavior of a user abruptly changes (increases or decreases). The filter will only adapt to this change after a certain amount of time. During adaption it will naturally deviate from the current process.

The resulting formula that needs to be minimized considering the offset of the filter response are then

$$AIC(f) = 2k_{avg} + n \cdot \ln\left(MSE_{o,global}\right) \text{ and}$$

$$BIC(f) = k_{avg} \cdot \ln\left(n\right) + n \cdot \left(MSE_{o,global}\right)$$

with respect to

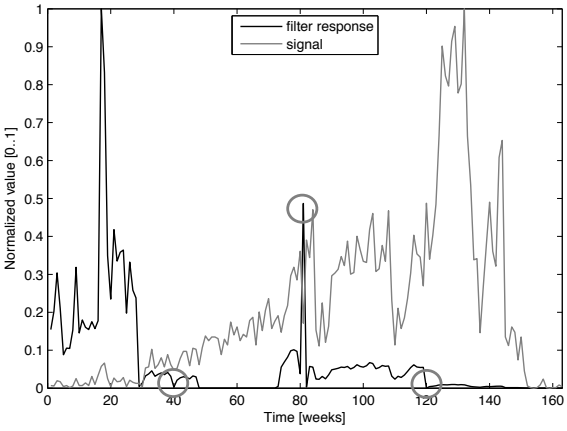


Fig. 5 Artifacts resulting from change in set of coefficients.

$$MSE_{o,global}(f,X) = \frac{1}{\sum\limits_{x\in X}\|x\|} \cdot \sum_{x\in X} \sum_{i=1}^{\|x\|-o(f)} (x_i - y(i+o(f)))^2.$$

2.5 Improvements

The problems arising from drastically changing user behavior were already discussed before and led to the offset being incorporated into the objective function. Another approach here could be an adaptive filter, that changes its passband frequency according to the history of latest events to lower the overall error even further. The problem arising from this approach occurs when using a high resolution for the time scale. Ultimately, almost all bins will be empty with only a few bins containing a single event. If a filter were to adapt to only a short history (e.g. five time bins) it would result in a constant filter response, namely a constant 0 signal, which reduces the MSE to almost 0.

But even using a lower resolution on the time scale produces artifacts arising from the fact, that changes in the passband filter may change single coefficients by orders of magnitude, leading to a totally different interpretation of the stored historical values in y. The resulting artifacts are shown in Fig. 5. Here the best fitting frequency for the first few weeks was calculated and then recalculated based on the whole previous signal every time the error value exceeded a given threshold (compared to the previous error). The following filter responses were calculated based on the new coefficients and the old signal, thus imitating a system, where new events are fed into the graph

and only previous information can be accessed. The red-circled points show some extreme cases where the change of coefficients led an abrupt change in the filter response without any evidence in the original data that supports these changes. Especially in the 120^{th} week the filter response drops to nearly zero although there is a drastic increase in activity in the original signal. Between week 50 and week 70 the filter response is almost constantly zero due to the poor parameters chosen at week 40.

A system that continuously changes the frequency on a low-to-medium time resolution, i.e. larger bins, in only small steps, thus gradually adapting the frequency, may lead to better results here.

2.6 Summary

Applying a Butterworth filter can be used to describe event frequencies in event-based graphs as continuous signal as opposed to the inherent discrete nature of the signal. The resulting curve is continuous, smooth and without overfitting it gives a generally good approximation of the original signal. The filter itself can be described only by its coefficients and a few historical entries for each edge based on the grade of the filter, leading to an overall efficient memory usage. Still this approach leaves enough space for adjustments, e.g. by weighting the extrema or the MSE differently in the objective functions and thus leading to curves being smoother or closer to the discrete signal. When changing the grade of the filter an even better approximation of the original curve is possible, decreasing the overall error at the cost of memory efficiency.

3 Analysis of Functional Connectivity Networks

In the last decade, a new trend in neuroscience emerged which focuses on the analysis of complex brain networks (see e.g. [31, 28]). These networks are commonly obtained from neuroimaging data coming from, e.g. electroencephalography (EEG), electrocorticography (ECoG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI). They measure some kind of brain activity of a certain region (e.g. on the skull, on the brain meninges, inside of the brain). We will call the brain regions that are to be measured *variables*.

Whenever two brain regions are co-active in some sense, we say that these regions are connected with each other. These connections induce a complex brain network that serves as some high-level representation of the underlying neural connectivity patterns. Certainly, it seems natural to analyze the functional brain connectivity using graph-based approaches. However, there are

many challenging research problems that have been and will be tackled when converting neuroimages to networks [4, 32]. Two fundamental questions which we are not going to address here relate to the definitions of both vertices and edges in the brain network.

In this paper the choice of vertices and edges is answered explicitly by the application we consider. Vertices are EEG channel positions on the skull. Edges correspond to the pairwise signal similarity between verticies. To be precise, we deal with the analysis of dynamic brain networks induced by patients' EEG. These patients have some special kind of visual field deficit, i.e. optic nerve damages [37]. We want to find network features that correlate with certain clinically relevant variables. Since we deal with (partly) blind subjects, we assume that their impairment has a major effect on the whole functional connectivity of each subject. Furthermore we hypothesize that there might be a correspondence between the degree of the visual loss and certain network metrics.

It has been shown already that brain damages include significant and longlasting neurological deficits [28]. So, a structural network disorder causes a functional network damage that might be observable by neuroimaging methods. The first study related to this was performed by the group of Cornelius Stam in 2007 [29]. This group analyzed the differences in EEG data between 15 patients with Alzheimer's disease (AD) and 13 control subjects. Functional connectivity was computed using synchronization likelihood (SL) [30]. The obtained brain networks have been measured by small-world network criteria [34]. Correlating these measures with clinical variables, they could show that AD might be characterized by a loss of small-world network characteristics. Based on these findings, we want to apply this idea to patients suffering from vision loss.

3.1 Functional Connectivity

As we already mentioned, it is necessary to define and measure functional connectivity between brain regions in order to obtain a complex brain network from neuroimaging data. Note that estimating functional connectivity does not necessarily mean finding the causal connections of the human brain. Functional connectivity can thus be only interpreted as a statistical relationship between brain regions. There might not be any causal coherence [23]. Nevertheless, a huge variety of different functional connectivity methods can been found in the literature. Here, we only want to mention a couple of them which are either related to or used in our work. For a deeper overview, see for instance [35].

A very common approach for EEG data is to use multivariate autoregressive (MVAR) models, e.g. directed transfer function (DTF) [12], Granger causality [8, 26]. MVAR models have been exhaustively used in the men-

tioned studies to find the signal frequencies of two brain regions that linearly correlate with each other. Since MVAR models measure the similarities of time series linearly, non-linear relations cannot be detected but might still be present in brain activity [30]. MVAR models are not well-suited to cope with spontaneous brain activity since they assume that the underlying process is stationary at any point in time[21, 22].

Therefore different non-linear methods have been proposed to measure the synchronization of two time series [35]. We just mention the synchronization likelihood (SL) [30, 20]. Consider a multivariate time series such as a multi-channel EEG recording of length N with n channels (the variables). We say that the measurement $x_{i,k}$ has been observed at timestamp i in channel k. First, a time-delay embedding is computed by

$$X_{i,k} = \left(x_{i,k}, x_{i+L,k}, x_{i+2\cdot L,k}, \ldots, x_{i+(m-1)\cdot L,k}\right)$$

where L is the lag and m the dimension of the embedding. These state vectors $X_{i,k}$ shall capture the relevant patterns of the signal. If we now consider only two channels A, B, then we can define a probability that $X_{i,k}$ are closer to each other than ε

$$P^{\varepsilon}_{i,k} = \frac{1}{2(W_2 - W_1)} \sum_{\substack{j \\ W_1 < |i-j| < W_2}}^{N} \theta(\varepsilon - d(X_{i,k}, X_{j,k}))$$

where d is typically the Euclidean distance. For each k and i the so-called critical distance $\varepsilon_{i,k}$ can be computed such that $P^{\varepsilon_{i,k}}_{i,k} = p_{\text{ref}}$ whereas $p_{\text{ref}} \ll 1$ is some user-defined threshold. Then for each pair of points in time (i, j) within $W_1 < |i-j| < W_2$, the number of channels $H_{i,j}$ for which $d(X_{i,k}, X_{j,k}) < \varepsilon_{i,k}$ is computed by

$$H_{i,j} = \theta(\varepsilon_{i,A} - d(X_{i,A}, X_{j,A})) + \theta(\varepsilon_{i,B} - d(X_{i,B}, X_{j,B}))$$

where θ is the Heaviside step function, $\theta(x) = 0$ if $x \leq 0$ and $\theta(x) = 1$ for $x > 0$. The synchronization likelihood is then given by

$$SL_i = \frac{1}{2p_{\text{ref}}(W_2 - W1)} \sum_{\substack{j \\ W_1 < |i-j| < W_2}}^{N} (H_{i,j} - 1) \tag{1}$$

Note the large set of free SL parameters, i.e. lag L, dimension m of the embedding, Theiler correction window W_1, window length W_2, reference probability p_{ref} with $0.01 \leq p_{\text{ref}} \ll 1$. Using prior information about the frequency range and temporal resolution of the signal [20], these 5 parameters can be reduced to p_{ref} only. When low-pass and/or band-pass filtering EEG, this information is given before computing SL.

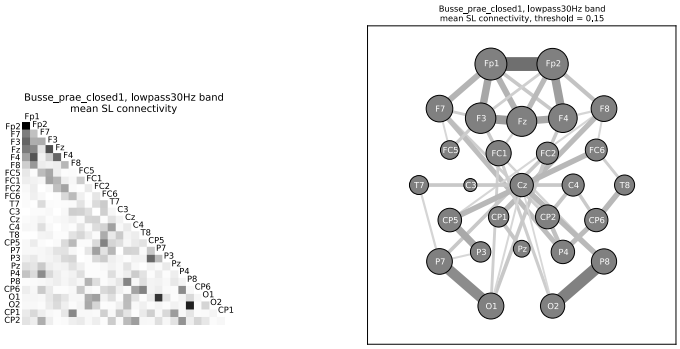


Fig. 6 Two visualizations of averaged functional EEG connectivity from one subject

3.2 Brain Graphs

A brain graph is created when computing functional connectivity, e.g. using SL, for each pair of variables at a given point in time. Such a network simply serves as graphical representation of pairwise statistical dependencies among all variables. For simplicity we demand that brain graphs are simple, i.e. they do not have any loops or multiple edges. Furthermore, since dealing with SL, we know that SL brain graphs must be symmetric.

Fig. 6 shows such a brain graph. Here, the averaged functional EEG connectivity from one subject has been computed by pairwise evaluation of SL. On the left hand side, only the lower half of the adjacency matrix is shown which is due to SL's symmetry. This matrix has been thresholded to contruct the graph on the right hand side. The threshold of 0.15 has been chosen carefully to diminish very weak connections while still preserving many of them. The size of each vertex corresponds to its weighted degree δ. With regard to reproducibility, the brain graphs can look completely different from patient to patient. This unfortunate fact is due to both corrupted EEG that has been drastically filtered and patients' differences in resting-state EEG.

3.2.1 The Meaning of Edges

An edge represents some kind of statistical dependency between two brain regions, i.e. the functional connectivity as explained in Sect. 3.1. The edge weight corresponds to the strength of the respective functional connectivity. Most similarity measures are normed to $[0, 1]$ or $[-1, 1]$ which enables a straightforward interpretation of the value of an edge weight. Commonly, researchers do not use weighted edges for graph analysis. Instead a (most often arbitrarily) chosen threshold is used to binarize the brain graph. Despite

the loss of information, some researchers argue that one can show different effects with a binary graph [25].

3.3 Experiments

In our experiments we used EEG data from 24 visually impaired subjects suffering from optic nerve damages [37]. In order to be able to relate EEG graph measures to clinical variables, so-called visual field charts were obtained from every patient. They indicate the location and size of the optic nerve damage. An expert defined 6 clinical measures based on the visual field charts.

To preprocess the EEG data we applied the following steps in EEGLAB [7]:

- manually removal of noisy time frames at beginning/end of each recording,
- removal of uncommon EEG channels across all subjects (28 were used),
- high-pass filtering with cutoff frequency at 1 Hz to remove slow movements,
- notch filtering 50 Hz and its harmonics up to 250 Hz to cope with European power line frequency,
- re-referencing by the average electrode,
- down-sampling to 250 Hz to reduce the costs of SL computation,
- linear trend removal,
- removal of biological artifacts using independent component analysis [18].

These artifacts that stem from electromyographic (EMG) or electrocardiograph (EKG) signal appear as noise in the recorded EEG signal in all variations. For EMG/ECG removal, ICA was applied to very carefully remove noisy components.

We used filters to obtain the conventional separation of frequency bands, since they are typically associated with different brain states [15, 9]. These frequency bands are δ: $f \in (1, 4]$ Hz, θ: $f \in (4, 8)$ Hz, α: $f \in [8, 13]$ Hz, β: $f \in (13, 30]$ Hz, γ: $f \in (30, 100]$ Hz. Furthermore we applied a broadband lowpass filter, i.e. $f \in (1, 30]$. Functional connectivity was established by SL [30]. The problem of choosing the SL parameters $W_1, W_2, L, m, n_{\mathrm{rec}}, p_{\mathrm{ref}}$ has been reduced [20] to two parameters, i.e. $n_{\mathrm{rec}} = 10$, $p_{\mathrm{ref}} = 0.01$. Every 10th point in time (corresponding to a frequency of 25 Hz), a new SL matrix was computed. We averaged the resulting series of graphs.

After obtaining the mean graphs for each subject and frequency band, we computed the following graph measures: average shell index, average path length, assortativity, average Kleinberg's authority, number of motifs of size 4, independence number, average eigenvector centralities, number of edges, density, average closeness centrality, local efficiency, number of motifs of size 3, clique number, average geodesic length in components, vertex connectivity, global efficiency, average betweenness centrality, average clustering coefficient, edge connectivity, diameter or longest geodesic, average eccentricity [5].

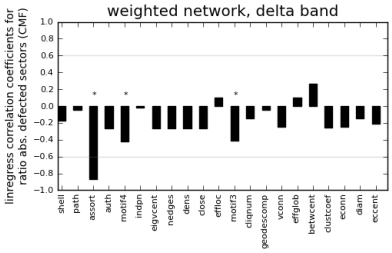


Fig. 7 Linear regression coefficients of weighted networks in the δ band related to the proportion of absolutely defected sensors using CMF.

3.4 Correlation Analysis

In order to test the strength of the above mentioned measures applied to the computed graphs, we correlated each clinical variable by computing the linear correlation coefficient. Two-sided p-values have been computed for hypothesis tests. Here, the null hypothesis is that the slope of the linear regression line is zero. This correlation analysis was applied to both the collection of weighted graphs and for the set of binary graphs by thresholding edge weights. Due to mostly weak edge weights close to zero, thresholds of 0.1 and 0.2 were applied. In the following we assume that a correlation is significant if its coefficient is equal to or higher than 0.6. Furthermore we marked correlations with an asterisk whenever the probability to reject the null hypothesis less than 5 percent.

For the weighted networks we have found only very few significant correlations. Some of them are situated in the δ-band when looking at the proportion of absolutely defected sensors using the Cortical Magnification Factor (CMF) [6] (see Fig. 7). Especially the density-based measures (i.e. number of motifs of size 4, number of edges, density, local efficiency, number of motifs of size 3, global efficiency, edge connectivity) correlate negatively with this clinical variable. This meets the intuition as a bigger visual deficit should have a negative effect on functional connectivity [28, 29].

All other significant correlations from the weighted networks have been found in the γ band for the reaction time. This is shown in Fig. 8. The basic hypothesis here is that a higher reaction time corresponds to a less efficient network due to higher path lengths.

For the simple graphs (i.e. binarized and thus without weights), we could only find some rather weak correlations for both thresholds in the γ band. Fig. 9 states that the higher the proportion of relatively defected sectors, the smaller the clustering coefficient. So, these networks show more random connectivity than the ones from subjects with smaller proportion of relatively defected sectors.

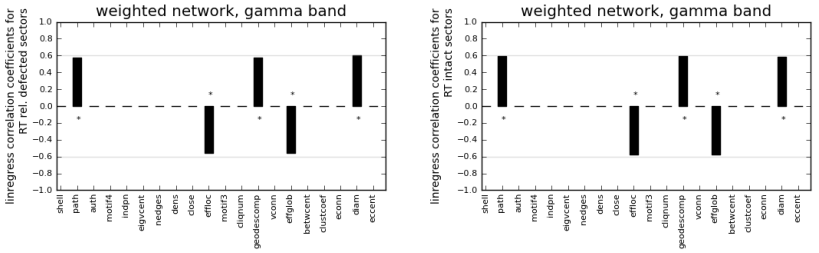


Fig. 8 Linear regression coefficients of weighted networks in the γ band showing less efficiency with increasing reaction time.

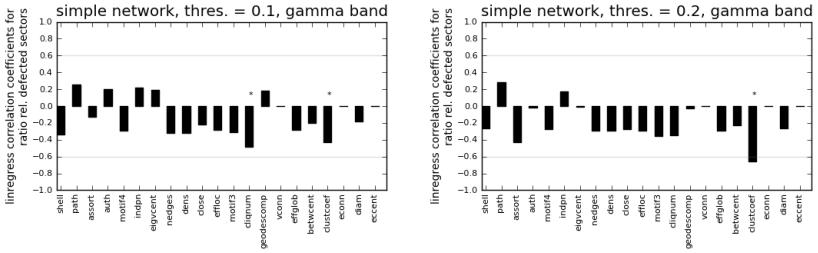


Fig. 9 Linear regression coefficients of simple networks in the γ band showing smaller clustering coefficients with a higher proportion of relatively defected sectors.

3.5 Summary

Until now, it is still unknown whether EEG features can properly describe damages of the human visual system. The goal of this study was to find both suitable network measures and useful clinical features that can be used for further patients analyses.

We therefore studied dynamic functional networks from patients with visual field defects. Based on typical frequency bands, the networks have been created by applying synchronization likelihood to several EEG time series from the subjects. We averaged the resulting series of graphs. Every averaged graph was described by several graph measures. The measures have finally been correlated to certain clinical variables describing each patient. In only some frequency bands we have found few significant correlations for a couple of variables and network measures. The correlations show that most prominently the γ band seems to be a fair marker for the effects of optic nerve damages. Also, the proportion of relatively defected sectors turned out to be the most informative clinical variable in this analysis.

4 Conclusions

We investigated two complex network problems demanding hybrid analysis methods from both intelligent data analysis and network theory. We dealt with the analysis of dynamic graphs from social science and neuroscience. Edge weights have been used in both types of networks to answer distinct questions. Firstly, we proposed a method to efficiently represent the strength of a relation between two entities based on events involving both entities. Using the Butterworth filter we were able to establish a continuous series of edge weights and thus graphs. Secondly, we analyzed EEG data of patients suffering from optic nerve damage. We showed how functional brain networks can be obtained and measured. The new measures have been correlated with clinical variables to find features describing the vision loss based on EEG. These features will be used in future work to guide the way to a clinical decision support system.

References

1. Alberts D, Cattaneo G, Italiano GF (1997) An empirical study of dynamic graph algorithms. *J Exp Algoritm* 2
2. Berthold MR, Borgelt C, Höppner F, Klawonn F (2010) Guide to Intelligent Data Analysis: *How to Intelligently Make Sense of Real Data*. Springer-Verlag, London
3. Butterworth S (1930) On the theory of filter amplifiers. *Exp Wirel & Wirel Eng* 7:536–541
4. Butts CT (2009) Revisiting the foundations of network analysis. *Science* 325(5939):414–416
5. Csárdi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal Complex Systems* 1695
6. Daniel PM, Whitteridge D (1961) The representation of the visual field on the cerebral cortex in monkeys. *J Physiol* 159(2):203–221
7. Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134(1):9–21
8. Ding M, Chen Y, Bressler SL (2006) Granger causality: Basic theory and application to neuroscience. In: Schelter B, Winterhalder M, Timmer J (eds.) *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, 437–460. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
9. Engel AK, Fries P, Singer W (2001) Dynamic predictions: Oscillations and synchrony in top-down processing. *Nat Rev Neurosci* 2(10):704–716
10. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. *Proc. Conf. on applications, technologies, architectures, and protocols for computer communication, SIGCOMM'99*, 251–262. ACM Press, New York
11. Fischhoff IR, Sundaresan SR, Cordingley J, Larkin HM, Sellier M, Rubenstein DI (2007) Social relationships and reproductive state influence leadership roles in movements of plains zebra, equus burchellii. *Anim Behav* 73(5):825–831
12. Kaminski MJ, Blinowska KJ (1991) A new method of the description of the information flow in the brain structures. *Biol Cybern* 65(3):203–210
13. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680

14. Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models, and methods. *Proc. 5th Int. Conf. on Computing and Combinatorics COCOON'99*, 1–17. Springer-Verlag, Berlin Heidelberg New York
15. Klimesch W (1999) EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res Rev* 29(2–3):169–195
16. Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD'06*, 611–617. ACM Press, New York
17. Lahiri M, Berger-Wolf TY (2010) Periodic subgraph mining in dynamic networks. *Knowl Inf Syst* 24:467–497
18. Makeig S, Bell AJ, Jung T, Sejnowski TJ (1996) Independent component analysis of electroencephalographic data. In: Touretzky DS, Mozer MC, Hasselmo ME (eds.) *Advances in Neural Information Processing Systems*, 8:145–151. MIT Press, Cambridge
19. Moewes C, Bola M, Sabel BA, Kruse R (2011) Brain connectivity associated with different damages of the visual system. *NIPS 2011 Satellite Meeting on Causal Graphs: Linking Brain Structure to Function*
20. Montez T, Linkenkaer-Hansen K, van Dijk BW, Stam CJ (2006) Synchronization likelihood with explicit time-frequency priors. *Neuroimage* 33(4):1117–1125
21. Nolte G, Bai O, Wheaton L, Mari Z, Vorbach S, Hallett M (2004) Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clin Neurophysiol* 115(10):2292–2307
22. Nolte G, Ziehe A, Krämer N, Popescu F, Müller K (2010) Comparison of granger causality and phase slope index. *J Mach Learn Res, Workshop and Conference Proceedings, Causality: Objectives and Assessment* 6:267–276
23. Pearl J (2009) Causal inference in statistics: An overview. *Stat Surv* 3:96–146
24. Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. *Proteins: Struct Funct Bioinf* 54(1):49–57
25. Rubinov M, Sporns O (2010) Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage* 52(3):1059–1069
26. Seth AK (2010) A MATLAB toolbox for granger causal connectivity analysis. *J Neurosci Methods* 186(2):262–273
27. Snyman JA (2005) Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms. *Applied Optimization*, vol. 97. Springer Science+Business Media, Inc., New York
28. Sporns O (2010) Networks of the Brain. MIT Press, Cambridge
29. Stam CJ, Jones B, Nolte G, Breakspear M, Scheltens P (2007) Small-World networks and functional connectivity in Alzheimer's disease. *Cerebral Cortex* 17(1):92–99
30. Stam CJ, van Dijk, BW (2002) Synchronization likelihood: an unbiased measure of generalized synchronization in multivariate data sets. *J Phys D: Nonlinear Phenom* 163(3–4):236–251
31. Varela F, Lachaux J, Rodriguez E, Martinerie J (2001) The brainweb: phase synchronization and large-scale integration. *Nat Rev Neurosci* 2(4):229–239
32. Wang J, Wang L, Zang Y, Yang H, Tang H, Gong Q, Chen Z, Zhu C, He Y (2009) Parcellation-dependent small-world brain functional networks: A resting-state fMRI study. *Hum Brain Mapp* 30(5):1511–1523
33. Wassermann S, Faust K (1997) Social Network Analysis: Methods and Applications. *Stuctural Analysis in the Social Sciences*, vol. 8. Cambridge University Press, Cambridge
34. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442
35. Wendling F, Ansari-Asl K, Bartolomei F, Senhadji L (2009) From EEG signals to brain connectivity: A model-based evaluation of interdependence measures. *J Neurosci Methods* 183(1):9–18

36. White DR, Harary F (2001) The cohesiveness of blocks in social networks: Node connectivity and conditional density. *Sociol Methodol* 31(1):305–359
37. Wüst S, Kasten E, Sabel BA (2002) Blindsight after optic nerve injury indicates functionality of spared fibers. *J Cogn Neurosci* 14(2):243–253
38. Zhang J (2010) A survey on streaming algorithms for massive graphs. In: Aggrawal CC, Wang H (eds.) *Managing and Mining Graph Data, Advances in Database Systems*, 40:393–420. Springer Science+Business Media, LLC, New York

Fuzzy Hyperinference-Based Pattern Recognition

Mario Rosario Guarracino[1], Raimundas Jasinevicius[2], Radvile Krusinskiene[2], and Vytautas Petrauskas[2]

Abstract The paper presents a new approach to the problem of pattern recognition. First of all, here is emphasized that the problem itself is fuzzy enough. Later three following novelties of the approach are disclosed: 1) the rule-based fuzzy inference, concerning the measure of patterns' similarity, is enriched by an idea of hyperinference; 2) a description of the main pattern recognition process is based on Takagi-Sugeno (T-S) reasoning procedure and 3) rule weights in T-S procedure are defined, solving special linear or piecewise linear programming problem (LPP or PWLPP), constructed according to the certain fuzzy experts' information. The proposed approach was used successfully for recognition of healthy people and those who suffer from certain illness (for example, an atherosclerosis). The classification was performed according to person's clinical posturograms (stabilograms). At the end of this paper experimental results are presented as well as acknowledgement to all anonymous participants of the experiments.

1 Introduction

All theoretical and practical activity of human individuals and their communities is based on data analysis, information mining, defining and extraction of certain features, which describe various characters, images, patterns, processes and even laws, followed by their grouping, clustering and recognition. And only the act of recognition serves as a base for certain decision making and action. The pattern recognition problem is old enough, very well-known

[1] Institute for High Performance Computing and Networking - National Research Council of Italy, Via Pietro Castellino 111, 80131 Napoli, Italy, mario.guarracino@cnr.it
[2] Department of Informatics - Kaunas University of Technology, Studentu 50-204a, LT-51368 Kaunas, Lithuania, {raimundas.jasinevicius, radvile.krusinskiene, vytautas.petrauskas}@ktu.lt

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 223–239.
springerlink.com

and perfectly described in scientific literature. The [1-3] may be named as the highly recommended texts concerning this topic and including an extremely good list of references. In general, a formulation of this problem itself implies a pretty good amount of fuzziness. According to [2], pattern recognition is defined as a search for structure in data, which is performed in three steps: data acquisition, pattern feature extraction from data by dimensionality reduction and mapping of extracted features to pattern classes. The [4] represents the newest approach and technique to the same problem based on the regularization of a generalized eigenvalue classification. The mapping process can be realized mainly, using two types of paradigms. The first one (a paradigm of determinism) is based on an existence of a set of crisp and perfectly standardized instructions, describing each pattern. It means that a decision maker (human being or machine) knows in advance presence (or absence) of which features determines belonging of a set of features under consideration to a certain pattern. The second paradigm (a paradigm of uncertainty) is based on a fact that a decision maker knows in advance for sure only several sets of features (examples) belonging to this or that pattern. So say, these sets are labeled by patterns' names, and it is supposed that some fuzzy descriptions of each pattern (generalized pattern) can be constructed using different teaching or training procedures. This paper is constructed around the paradigm of uncertainty using the fuzzy inference approach. The approach is enriched by mechanism of hyperinference [5, 6], which is included into widely spread Takagi-Sugeno (T-S) reasoning procedure [2]. The main teaching or training process for parameters and weights determination in the T-S scheme is substituted by solving an adequate mathematical programming problem formulated according to the fuzzily described expert requirements [7, 8]. The paper consists of four main sections. Section 1 presents an idea of fuzzy rule-based hyperinference in pattern recognition. Section 2 - describes the main pattern recognition process based on Takagi-Sugeno (T-S) reasoning procedure. Section 3 is dedicated to special linear or piece-wise linear programming problems (LPP or PWLPP), constructed according to certain fuzzy experts' information to define all necessary rule weights in T-S procedure. Section 4 contains results of real experiments on the use of the proposed approach for recognition of healthy people and those who suffer from certain illness (in our case - an atherosclerosis). The paper ends with special thanks and acknowledges to all anonymous participants of those experiments as well as to all supporters of this research.

2 Fuzzy Rule-based Hyperinference for Pattern Recognition

Ordinary fuzzy systems inference is based: 1) on a deriving verbal (linguistic) or parametric consequents by preprocessing lists of fuzzy rules, containing

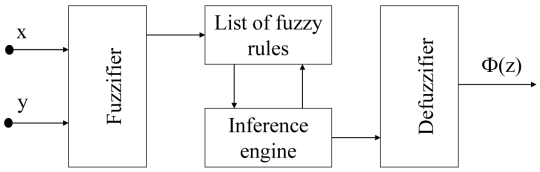


Fig. 1 A block diagram of an ordinary fuzzy system

verbal or parametric antecedents connected by certain fuzzy logic operations and 2) on a defuzzification process using some compositional rule or formula [2, 7]. Types of rules can be presented as follows:

IF x is A AND y is B THEN z is C (for Mamdani fuzzy models) (1)
IF x is A AND y is B THEN $z = F(x, y)$ (for Takagi-Sugeno fuzzy models)

Defuzzification procedures for the two cases mentioned above can be described as a reasoning on the base of a set of consequents C using the CoG (center of gravity), or MoM (mean of maximum) methods for Mamdani type systems [2, 7], and MF (fuzzy mean) method as a reasoning by evaluation of all results z included and processed according to the certain formula $\Phi(z)$ for Takagi-Sugeno systems. A block-diagram of an ordinary fuzzy system corresponding to both cases is presented in Fig. 1. As it is emphasized in different references, and especially in [5] and [6], all ordinary fuzzy systems process so-called *positive rules* which for cases expressed by (1) have general form:

IF <*condition*> THEN <*action or rating*> RECOMMENDED (2)

But the real life often requires taking into account various factors and conditions which have a meaning of warning, precaution and even prohibition. In such a case fuzzy system must process so-called *negative rules*:

IF <*condition*> THEN <*action or rating*> NOT RECOMMENDED (3)
(WARNED AGAINST/ PROHIBITED)

Almost each more complex and sophisticated real case under investigation is described by experts and decision makers using some set or mixture of positive and negative rules. The drawback of the ordinary fuzzy systems is their impossibility to cope with the mentioned mix of rules, which can be circumvent by fuzzy *hyperinference* process [5]. The hyperinference performs an evaluation of the influence of fuzzy consequents z derived from positive as well as negative rules. It requires different and more sophisticated fuzzy logic formulae and new and two-way fuzzy system structure (Fig. 2) described and delivered in [5].

As a matter of fact, a pattern recognition task, as presented in the introduction, belongs to the class of fuzzily described problems and looking

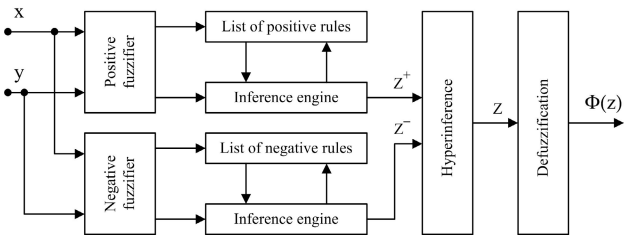


Fig. 2 A block diagram of two-way fuzzy hyperinference system

for a defuzzified answer. According to the paradigm of uncertainty, a decision maker knows in advance several sets of features (examples) belonging to one or another pattern. And a task is to use those feature sets, and to use some sets of fuzzy rules for Takagi-Sugeno reasoning procedure to construct a generalized pattern for each possible class. It is worth of mentioning that in [6] an idea of reasoning by using something similar to two types of rules (positive and negative ones) for pattern recognition was firstly proposed. Actually, the construction of generalized patterns was based on extraction and emphasizing of two types of features: those which are most common for the given pattern ("positive similarities" inside the class), and those which distinguish the pattern under investigation from all possible competing classes (dissimilarities between certain pattern and all other classes, or "negative similarities"). Such an approach requires some extension of the fuzzy logic inference procedure, which is called hyperinference. The next section of this paper is dedicated to elaborate the application of fuzzy hyperinference-based Takagi-Sugeno reasoning procedure for pattern recognition.

3 Pattern Recognition based on Takagi-Sugeno (T-S) Reasoning Procedure

If a pattern of a certain class is considered as a physical or abstract structure of class' objects, described by a set of distinctive features [1, 2], then a simplified pattern recognition problem can be formulated in the following way. Let us imagine our world under investigation as consisting of S classes of objects, where each class p has its own pattern. So, $p = 1, 2, \ldots, r, \ldots, S$. Each object is described by N features numbered as $n = 1, 2, \ldots, j, \ldots, N$. In a case when certain feature extraction, measurement and normalization procedures are performed [6], the i-th feature of an object, belonging to the p-th class (corresponding to the p-th pattern) can be represented by a real number α_{pi} which expresses a degree of intensity of this particular feature. It is convenient to use a vector-row notation to describe the whole object $\boldsymbol{\alpha}_p = (\alpha_{p1}, \ldots, \alpha_{p2}, \ldots, \alpha_{pi}, \ldots, \alpha_{pN})$. If we have several objects (their num-

bers are $l = 1, 2, \cdots, k, \cdots, L$) and know in advance that they belong to class p (they are originated by the p-th pattern), than we can say that the class p is represented by a set of vectors $\boldsymbol{\alpha}_p^l$ ($l = 1, 2, \cdots, k, \cdots, L$). The main task of pattern recognition procedure consists in developing several Takagi-Sugeno type rules and defuzzification instruments. This must be done: a) using whole available information about the patterns, which is hidden in the set of $\boldsymbol{\alpha}_p^l$, $\forall p$; and b) using whole available experts' experience, which is presented in verbal form and was collected working with objects' and patterns' features. A complex of such actions enables to construct a pattern recognition instrument capable to assign any unknown but properly described object $\mathbf{x}$ to one of the possible patterns (or classes): $1, 2, \ldots, r, \ldots, S$. The accuracy of this assignment depends on the instrument's decision making efficiency to process this fuzzy information. Usually better reasoning results are achieved when features of objects are not only normalized but centred as well [8, 9]. It means that the whole object is represented as a vector $\boldsymbol{\alpha}_p^{o^l} = (\alpha_{p1}^{o^l}, \ldots, \alpha_{p2}^{o^l}, \ldots, \alpha_{pi}^{o^l}, \ldots, \alpha_{pN}^{o^l})$ with components calculated according to the following formula:

$$\alpha_{pi}^{o^l} = \alpha_{pi}^l - \frac{1}{N}\sum_{j=1}^{N} \alpha_{pj}^l \tag{4}$$

In such a situation, the Takagi-Sugeno (T-S) reasoning procedure combined with the hyperinference process can be built using concepts represented by (1)-(3). It means that a set of "positive rules" for (T-S) pattern recognition procedure consists of a list of statements such as:

IF <degree of certainty, that feature i with intensity x_i^o belongs to the pattern p, is K_{pi} > THEN $< z_i^+ = K_{pi}x_i^o >$ RECOMMENDED

The latter can be written as:

$$\text{IF} < \mu^+(x_i^o) = K_{pi} > \text{THEN } z_i^+ = K_{pi}x_i^o \quad \forall p, i \tag{5}$$

Similarly, a set of "negative rules" for (T-S) a pattern recognition procedure consists of a list of statements:

IF < degree of certainty, that feature i with intensity x_i^o belongs to any other pattern except p, is K_{pi} > THEN $< z_i^- = -K_{pi}x_i^o >$ NOT RECOMMENDED

Equivalently:

$$\text{IF} < \mu^-(x_i^o) = K_{pi} > \text{THEN } z_i^- = -K_{pi}x_i^o \quad \forall p, i \tag{6}$$

According to the concept of hyperinference [5]:

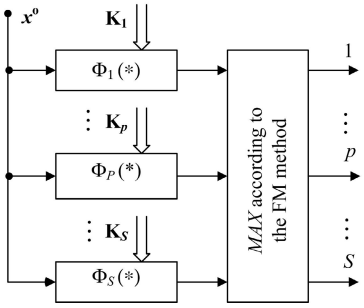


Fig. 3 Decision making act for T-S reasoning procedure

$$\mu(x_i^o) = max\{\mu^+(x_i^o), (1 - \mu^-(x_i^o))\} \tag{7}$$

and

$$z_i = max(z_i^+, z_i^-) = K_{pi} x_i^o, \forall p, i \tag{8}$$

When the whole unknown object $\mathbf{x}^o$ is under consideration, its degree of belonging to the pattern p can be evaluated by

$$\Phi_p(\mathbf{x^o}) = \sum_{i=1}^{N} K_{pi} \quad \forall p \tag{9}$$

This defuzzification method for (T-S) procedure is called the fuzzy mean (FM) and can be expressed in the vector notation form as

$$\Phi_p(\mathbf{x^o}) = \mathbf{x^o} K_p^T \quad \forall p \tag{10}$$

Here T is the transposition to row vector of:

$$\mathbf{K}_p = (K_{p1}, K_{p2}, \ldots, K_{pi}, \ldots, K_{pN}). \tag{11}$$

A block diagram, representing a final decision making act in the case of fuzzy pattern recognition based on Takagi-Sugeno (T-S) reasoning procedure, is shown in Fig. 3.

In practice, the possibility to verbally formulate understandable lists of rules for sets presented by (5) and (6) often does not exist. Usually neural-type training procedures based on gradient methods are used [1, 7]. In the next section of this paper a special linear or piece-wise linear programming problem (LPP or PWLPP) is proposed to find optimal values of those degrees of certainties K_{pi} for (5)-(8) and $\forall p, i$. LPP and PWLPP problems are formulated and constructed according to the certain fuzzy experts' information enabling us to define all necessary rule weights in the T-S pattern recognition procedure.

4 Linear and Piece-wise Linear Programming Problems for T-S Procedure Rule Weights

To avoid a comparatively unpredictable and very clumsy neural-type training procedures for determination of K_{pi} values $\forall p, i$ in the Takagi-Sugeno (T-S) reasoning procedure for pattern recognition, we formulate an objective function maximization problem subjected to a set of constrains, constructed according to a certain fuzzy experts' information. The structure of (9) and (10) implies a very simple (linear) form of a function to be maximized as well as linearity of constraints.

According to the pattern recognition problem's description described in the sections 2 and 3 of this paper, whole available information about the patterns is hidden in the set of $\boldsymbol{\alpha}_p^{o^l}$, $\forall p$, where $p = 1, 2, \ldots, r, \ldots, S$ and $l = 1, 2, \ldots, k, \ldots, L$. In spite of its fuzziness, a wise enough formulation of a problem for the determination of degrees of certainties K_{pi}, $\forall i$ in the Takagi-Sugeno (T-S) reasoning for recognition of a pattern for objects belonging to the pattern p can be constructed as follows. Let us select at random one representative of the class p , for example $\boldsymbol{\alpha}_p^{o^k}$ (and call it "central" only for a simplicity of understanding), and require to find such K_{pi}, $\forall i$ that the measure of degree of a certainty $\Phi_p(\boldsymbol{\alpha}_{\mathbf{p}}^{\mathbf{o^k}})$ of belonging of the selected object k to the pattern p would be maximum:

$$\Phi_p(\boldsymbol{\alpha}_{\mathbf{p}}^{\mathbf{o^k}}) = \sum_{i=1}^{N} \alpha_{pi}^{o^k} K_{pi} \rightarrow max \tag{12}$$

and it must be reached under following constrains:

$$\sum_{i=1}^{N} \alpha_{pi}^{o^l} K_{pi} \geq \gamma \sum_{i=1}^{N} \alpha_{pi}^{o^k} K_{pi} \quad \forall l. \tag{13}$$

and

$$\sum_{i=1}^{N} \alpha_{ri}^{o^l} K_{pi} \leq \kappa \sum_{i=1}^{N} \alpha_{pi}^{o^k} K_{pi} \quad \forall r \neq p, \ \forall l. \tag{14}$$

Optimal values of γ and κ are recommended from the interval $[0 - 1]$, and $\gamma > \kappa$ [8]. Concrete values of those coefficients depend on the experts' knowledge or guess concerning the patterns (or classes) structure (internal connections and dispersion of patterns' features). Physical meaning of the (13) is tightly connected with the understanding of "positive similarities" inside the class, and extracting of those similarities using the set of "positive rules", as it was mentioned in the section 2. Physical meaning of (14) corresponds to the concept of dissimilarities between certain pattern (in our case pattern of a class p) and all other classes $r \neq p$ (or "negative similarities"), and to the process of extracting of those dissimilarities by the set of "negative rules".

Coefficients γ and κ permit us to control the fuzzy level of those similarities and dissimilarities, and the solution of (12)-(14) implements the concept of fuzzy hyperinference. By the way, even fast investigation of the problem described above shows that the problem belongs to the class of linear programming problem (LPP) where inequalities (13) and (14) need additional constrains:

$$0 \leq \mathbf{K}_p \leq A, \tag{15}$$

where A is any practically convenient real number. Naturally, a solution of the LPP (12)-(15) for the pattern (class) p consists of the obtained value for

$$max\Phi_p(\boldsymbol{\alpha}_p^k) = \Phi_{max} \quad and \quad \mathbf{K}_p = (K_{p1}, K_{p2}, \ldots, K_{pi}, \ldots, K_{pN}). \tag{16}$$

The procedure must be repeated for all patterns p. In this way, the set of S solutions will be obtained, and the recognition procedure must be performed according to Fig. 3, taking into account the need of fulfilling proportionality condition:

$$c_1\Phi_{1\ max} = \cdots = c_p\Phi_{p\ max} = \cdots = c_S\Phi_{S\ max} = B; \tag{17}$$

where B and c_p are real numbers. This condition plays a role in the normalization procedure, and it means that the fuzzy (verbal) term "VERY SIMILAR" must be evaluated by the same number B, whichever pattern we are taking into consideration. Sometimes experts and decision makers have in advance an additional information concerning the internal structure of classes under consideration (for example, they guess an existence of certain subclasses in each pattern or so on). Let us assume that the pattern p consists of wp subclasses numbered as $w = w_1, w_2, \ldots, w_p$. Then, the classical linear programming problem (LPP) for determination of weights in the Takagi-Sugeno (T-S) reasoning procedure for pattern recognition can be substituted by a piece-wise linear programming problem (PWLPP) similar to the (12)-(15) [8]. The requirement is to maximize:

$$\Phi_p(\boldsymbol{\alpha}_p^{o^k}) = \max_w \left(\sum_{i=1}^{N} \alpha_{pi}^{o^k} K_{pi}^w \right) \rightarrow max, \tag{18}$$

where $w = w_1, w_2, \ldots, w_p$, under the constrains:

$$\max_w \left(\sum_{i=1}^{N} \alpha_{pi}^{o^l} K_{pi}^w \right) \geq \gamma \max_w \left(\sum_{i=1}^{N} \alpha_{pi}^{o^k} K_{pi}^w \right) \quad \forall l \tag{19}$$

$$\max_w \left(\sum_{i=1}^{N} \alpha_{ri}^{o^l} K_{pi}^w \right) \leq \kappa \max_w \left(\sum_{i=1}^{N} \alpha_{pi}^{o^k} K_{pi}^w \right) \quad \forall r \neq p, \forall l \tag{20}$$

$$0 \leq \mathbf{K}_p^w \leq A \quad \forall w \tag{21}$$

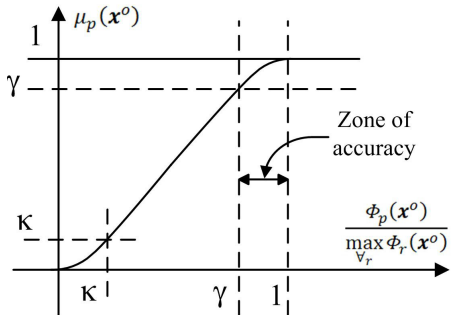


Fig. 4 Recommended certainty that object belongs to class p

The procedure must be repeated for all patterns p, and the condition (17) must be fulfilled as well. The PWLPP approach enables us to use more sophisticated Takagi-Sugeno (T-S) hyperinferrence reasoning procedure for pattern recognition problems solution based on MoM (mean of maximum) defuzzification method. Worth of mentioning the fact that in some cases the general PWLPP can be divided into several simple LPP according to the information about the patterns available in advance. It is clear that a maximal number of simple LPPs is $v = \sum_{p=1}^{S} w_p$. All intermediate cases between PWLPP and LPP correspond to different fuzzy logic operations involved in Takagi-Sugeno (T-S) hyperinference procedures. Some of them are demonstrated in the next section of this paper. As it was delivered earlier, the Takagi-Sugeno (T-S) hyperinference reasoning procedure supplies us with the degrees of certainty, that the description of an unknown object belongs to the pattern p for $\forall p$, as depicted in Fig. 3. If we want to use this information as a recommendation for decision maker, the value

$$u_p = \frac{\Phi_p(\mathbf{x}^o)}{\max_r \Phi_r(\mathbf{x}^o)}$$

must be determined. In practice, an interval for u_p is $[0-1]$, with some zone of accuracy as it is shown in Fig. 4, where the degree of truth $\mu_p(\mathbf{x}^o)$ is presented.

In case when u_p is in this zone, special additional investigations of the object's properties are strongly recommended before the final decision is taken. In Fig. 5 an extended block diagram is shown where both stages for pattern recognition process are represented. The first stage of this process corresponds to the (T-S) hyperinference reasoning procedure, and the second one - to the procedure of decision making recommendations.

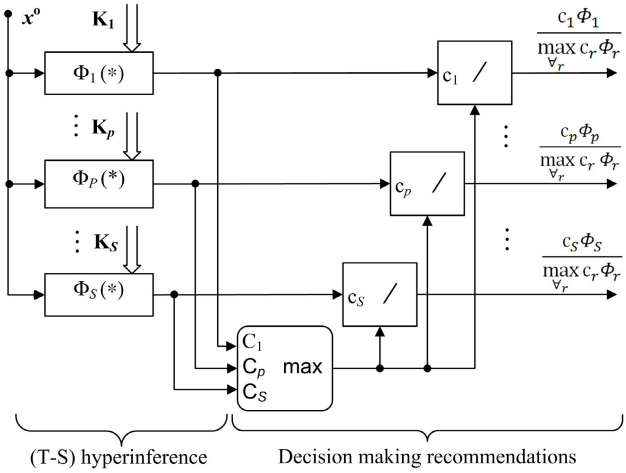


Fig. 5 (T-S) hyperinference and decision making recommendations procedure

5 Application of Hyperinference-based Pattern Recognition Procedures for Classification of Human Posture

The proposed approach was successfully used for recognition of strong (healthy) people and those who suffer from certain illness (for example, an arthrosclerosis). The classification was performed according to person's clinical posturograms (stabilograms). In this case, a data acquisition, pattern feature extraction from data and a dimensionality reduction were performed as described in the following.

5.1 Visualization of Human Posture Stability

One of the most popular ways to visualize standing stability is to register movements of centre of pressure (COP) of a human body on the base of the support [10, 11]. The resulting digitalized trajectory of the COP is referred to as a stabilogram. In general, the stabilogram is a collective outcome or result of activities of all systems that are responsible for maintaining the body in upright position. The typical stabilogram of a healthy subject on the base of support is presented in Fig. 6. All stabilograms were recorded during physical experiments.

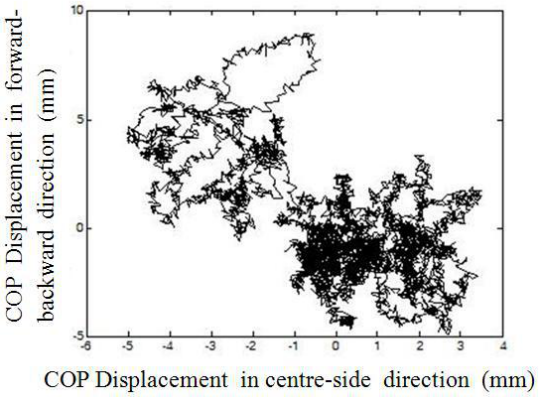


Fig. 6 Trajectory of COP movement on a base of support

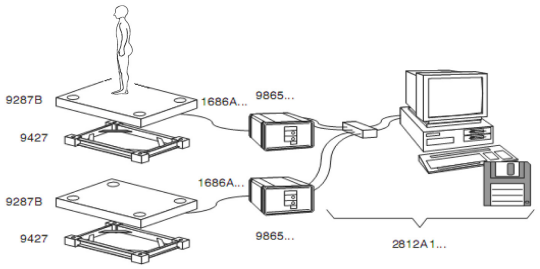


Fig. 7 Set up for clinical experiment

5.2 Protocol of Physical Experiments

Fifteen healthy (HL) and fifteen suffering from multiple sclerosis disability (MSD) subjects (all female) age 32±2 years (average standard deviation) took part in physical experiments. Subjects who suffer from multiple sclerosis were chosen to take part in experiments since this disease affects central neural system (CNS) of an individual and causes difficulties to maintain the stable bipedal posture [12]. The diagnosis of each participant was known in advance before the real experiments took place. The stabilogram was recorded for each subject. Participants were asked to stand on a base of support in a bipedal comfortable posture. The duration of the experiments was 60 sec. with a sampling rate of 100 Hz. Stabilograms were recorded using Kistler 9287B force platform (Fig. 7). All signals were centralized in space and in magnitude, i.e. the average position of COP signal has coordinates (0, 0). The planar (2D) histograms of stabilograms were calculated. Typical histograms of HL and MSD subject consisting of 25 bins (5 in Medio-Lateral multiplied by 5 in Anterio-Posteriori direction) are presented in Fig. 8. Bins are numbered as shown in the Fig. 8. The darker box represents more time of COP signal spent in the bin, i.e. the higher intensity of a feature.

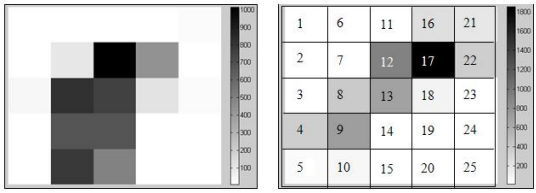


Fig. 8 Histogram of stabilogram: left - HL subject No. 15; right - MSD subject No. 22

5.3 Hyperinference-based Pattern Recognition Problem Formulation for Classification of Histograms of Human Posture

According to the hyperinference-based pattern recognition problem formulation presented in section 2, two classes of objects $\boldsymbol{\alpha}$ - histograms of HL individuals and histograms of MSD individuals - are investigated. It means that $S = 2$. Each object $\boldsymbol{\alpha}$ to be classified has $N = 25$ features - histogram bins. Order of their placing into object's description vector $\boldsymbol{\alpha}$ is shown in Fig. 8 (right histogram). Every object is centered according to (4). For those experiments "positive similarity" of objects was evaluated by constant $\gamma = 0.8$, and a degree of "negative similarity" was evaluated by constant $\kappa = 0.3$. The difference $(\gamma - \kappa)$ determines experts' fuzzy assumption concerning a possible structure of classes under investigation and was selected on the base of some experience [8].

All calculations of degrees of certainties K_{pi} $\forall p, i$ using experimental data were conducted with Matlab software, and LPPs (or PWLPP) were solved using its `linprog` function. Three experimental cases are delivered here.

A. All histograms were assigned to the HL or MSD class prior to experiment according to protocol of medical diagnosis and results of physical experiment. The aim of those calculations and experiment is to show that it is possible to identify hyperplane as a hypersurface distinguishing HL and MSD objects.
B. Several representatives were selected arbitrary to represent HL and MSD classes. Calculations of degrees of certainties K_{pi} for were conducted using LPP method.
C. The same as experiment B., but calculations were performed by means of PWLPP solving method.

In a numerical experiments' performance the procedure "arbitrary selected" was implemented by means of Matlab software function `rand`.

Table 1 Degrees of certainties K_{pi} for HL and MSD classes.

$\mathbf{K}_{\mathrm{HL}}$					$\mathbf{K}_{\mathrm{MSD}}$				
23.38	0.00	0.00	6.87	25.00	0.00	25.00	25.00	25.00	1.05
25.00	0.00	17.12	13.00	18.55	0.00	25.00	11.34	14.70	0.00
0.00	21.05	10.59	19.66	1.90	18.07	9.19	17.94	9.38	18.11
0.00	21.70	12.01	25.00	0.00	25.00	5.32	12.91	5.35	25.00
25.00	25.00	25.00	25.00	25.00	8.84	5.72	0.00	0.00	0.00

5.4 Results of A, B, C Experiments

Identification of a distinguishing hyperplane. As it was mentioned before in this experiment all histograms were assigned to HL or MSD class respectively. The aim of this experiment is to show that it is possible to classify these objects to two separate classes using (T-S) reasoning procedure when $\mathbf{K}_{\mathrm{HL}}$ and $\mathbf{K}_{\mathrm{MSD}}$ are determined as a LPP solution. As the "central" object representing HL class was selected object No. 14, and a "central" object No. 30 was selected for the MSD class. As a result of LPP, constructed according (12)-(15) for the first experiment (A.), the and vectors for HL and MSD classes were calculated, and they are presented in Table 1. For the (12)-(15) LPP an upper bound $A = 25$ (15) was selected only for practical convenience as a certain scale factor, and constants $\gamma = 0.8$ and $\kappa = 0.3$ were selected as fuzzy experts' recommendations enabling to evaluate positive and negative similarities between competing classes.

Results shown in Table 2 confirm that LPP was successfully solved and all histograms were correctly classified according to the decision making act for T-S reasoning procedure presented in Fig. 3. Results of experiment A confirmed the hypothesis, that subjects suffering form MSD have their own posture pattern and different than subjects from the HL class, and that this pattern may be: a) identified by means of (T-S) fuzzy inference and decision making procedure, and b) used for classification of posture patterns.

Arbitrary selected representatives for the LPP. As the "central" object representing HL class was arbitrary selected object No. 14, and together with other arbitrary selected objects (No. 3, 4, 7, 9 and 13) it constitutes a set of representatives for HL class. Similarly the MSD class has its own "central" object No. 30, and together with object No. 20 constitutes a set of representatives for MSD class. All other objects were set as "from unknown class and need to be classified". Vectors of features' significance $\mathbf{K}_{\mathrm{HL}}$ and $\mathbf{K}_{\mathrm{MSD}}$ of HL and MSD classes obtained after solving of the LPP (12)-(15) for this case are presented in Table 3. Table 4 shows the classification results. As it may be noticed classification procedure did come to wrong outcome in case of objects No. 2, 5, 6, 19, 21, 25 and 29.

Table 2 Classification of histograms according to T-S reasoning, when the class of all histograms was known in advance (experiment A.)

Similarity Function Φ	$\Phi_{\mathrm{HL}} = 10087.43$	$\Phi_{\mathrm{MSD}} = 8803.17$	Decision results
Normalization coefficient c_i	1	1.14	Decision rule
	$c_1\Phi_{\mathrm{HL}}$	$c_2\Phi_{\mathrm{MSD}}$	$\max(c_1\Phi_{\mathrm{HL}}, c_2\Phi_{\mathrm{MSD}})$
Histograms:			
HL class			
1	8069.94	3026.23	HL
2	12096.70	3026.23	HL
3	8069.94	3026.23	HL
4	17592.51	-13124.55	HL
5	11540.22	3026.23	HL
6	8069.94	3026.23	HL
7	19043.45	-9515.65	HL
8	8069.94	3026.23	HL
9	8069.94	3026.23	HL
10	13628.49	-7193.49	HL
11	21215.48	-11388.51	HL
12	35124.71	-26205.48	HL
13	8069.94	3026.23	HL
14	10087.43	-3463.13	HL
15	17779.39	-10462.85	HL
MSD class			
16	-12466.47	22391.81	MSD
17	-314.97	10602.41	MSD
18	3026.23	8069.94	MSD
19	-15149.95	24467.82	MSD
20	-2780.15	15780.06	MSD
21	3026.23	8069.94	MSD
22	1079.17	8069.94	MSD
23	-3892.17	19627.50	MSD
24	3026.23	9512.49	MSD
25	3026.23	8069.94	MSD
26	-9373.07	19614.88	MSD
27	-2244.26	11263.74	MSD
28	3026.23	8069.94	MSD
29	-2914.59	12705.14	MSD
30	2915.28	10087.43	MSD

Arbitrary selected representatives for the PWLPP. In this experiment objects representing HL class was arbitrary divided in three subclasses $w = w_1, w_2, w_3$ and were represented respectively:

- Subclass HL_1 was represented by objects No. 3 and 4 (object No. 4 was chosen as a "central" object of a given subclass);
- Subclass HL_2 was represented by objects No. 9 and 13 (object No. 13 was chosen as a "central" object of a given subclass);
- Subclass HL_3 was represented by objects No. 7 and 14 (object No. 14 was chosen as a "central" object of a given subclass).

Table 3 Degrees of certainties $\mathbf{K}_{\mathrm{HL}}$ and $\mathbf{K}_{\mathrm{HL}}$ for HL and MSD classes in case when HL class is represented by objects No. 3, 4, 7, 9, 13, 14 and MSD class is represented by objects No. 20, 30.

$\mathbf{K}_{\mathrm{HL}}$					$\mathbf{K}_{\mathrm{MSD}}$				
0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.74	0.00	25.00
0.00	17.73	0.00	13.56	0.00	25.00	25.00	25.00	25.00	25.00
0.00	0.00	17.56	7.84	0.00	0.00	25.00	12.99	25.00	0.00
0.00	25.00	25.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	25.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

MSD class was represented by a "central" object No. 30 together with object No. 20; they constituted a set of objects describing the MSD class. The PWLPP (18)-(21) was formulated and solved. As a solution the features' significance vectors $\mathbf{K}_{\mathrm{HL}}$ for HL subclasses HL_1, HL_2, HL_3 and $\mathbf{K}_{\mathrm{MSD}}$ for the class MSD class are presented in Table 5.

During the decision making stage the rest of objects was considered as "being from the unknown in advance class and needing to be classified." Table 6 shows the similarity functions and classification results received by the means of (T-S) fuzzy inference and decision making procedure according to the Fig. 5. It may be noticed that in this case the classification procedure gave nine wrong answers.

6 Concluding Remarks

In the theoretical part of this paper, a new approach to the problem of fuzzy pattern recognition was delivered. First of all the rule-based fuzzy inference, concerning the measure of patterns' similarity, was connected with the idea of hyperinference. Secondly a description of the main pattern recognition process was based on Takagi-Sugeno (T-S) reasoning procedure, enhanced by specific decision making recommendations. Rules' weights in (T-S) procedure are defined, solving special linear or piece-wise linear programming problem (LPP or PWLPP), constructed according to the certain fuzzy experts' information delivered or guessed in advance. The practical significance of the proposed approach was experimentally confirmed by using the proposed instrument for recognition and diagnostics of healthy people and those who suffer from an atherosclerosis. The classification was performed according to persons' clinical investigations data and real posturograms. According to authors knowledge such an approach was used for the first time in the clinical practice, and it opens a new chapter for further research in this field.

Acknowledgements First of all we express our thanks and appreciation to all those anonymous volunteers who took part in our clinical experiments. Secondly, we acknowledge the stimulating ideas to this investigation captured from prof. Christian Borgelt from the

Table 4 Histograms' classification according T-S reasoning, when HL class is represented by objects No. 3, 4, 7, 9, 13, 14 and MSD class is represented by objects No. 20, 30.

Similarity Function Φ	$\Phi_{\mathrm{HL}} = 55405.14$	$\Phi_{\mathrm{MSD}} = 67496.77$	Decision results
Normalization coefficient c_i	1	0.82	Decision rule
	$c_1\Phi_{\mathrm{HL}}$	$c_2\Phi_{\mathrm{MSD}}$	$\max(c_1\Phi_{\mathrm{HL}}, c_2\Phi_{\mathrm{MSD}})$
Histograms:			
HL class			
1	44323.69	16153.97	HL
2	15091.79	34814.82	MSD
3	44324.11	10938.26	HL
4	64094.18	3306.91	HL
5	-13217.47	54909.70	MSD
6	13170.49	54903.00	MSD
7	44324.11	16621.54	HL
8	34581.95	21986.19	HL
9	44324.11	7.68	HL
10	71746.21	-17248.80	HL
11	61086.85	-7648.34	HL
12	77497.88	-31443.36	HL
13	47943.41	16621.54	HL
14	55405.14	15.17	HL
15	62818.73	-9062.90	HL
MSD class			
16	3982.39	61184.81	MSD
17	31842.19	33519.37	MSD
18	28285.70	42313.67	MSD
19	40802.43	27715.89	HL
20	16621.54	52156.22	MSD
21	41255.05	28846.41	HL
22	14097.35	42526.12	MSD
23	754.36	56923.29	MSD
24	21122.40	48521.64	MSD
25	29314.99	20582.00	HL
26	20960.50	42191.91	MSD
27	17689.01	24905.22	MSD
28	18298.49	33782.13	MSD
29	49431.68	11767.80	HL
30	11516.27	55405.14	MSD
Number of errors			7

European Center for Soft Computing (in Mieres, Asturia, Spain). And thirdly, we would like to emphasize a positive influence of a financial support through COST program received from Lithuanian Agency for International Science and Technology Programs and COST IC0702 Action in general.

Table 5 Degrees of certainties K_{pi} vectors of subclasses for HL and MSD classes, when HL class was arbitrary divided to three subclasses.

$\mathbf{K}_{\mathrm{HL}}$					$\mathbf{K}_{\mathrm{MSD}}$				
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	25.00	0.00	0.00	25.00	25.00	22.59	0.00
0.00	0.00	13.23	25.00	0.00	0.00	25.00	25.00	25.00	0.00
0.00	25.00	25.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	22.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

$\mathbf{K}_{\mathrm{HL}}$					$\mathbf{K}_{\mathrm{MSD}}$				
0.00	1.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	25.00	25.00	25.00	0.00
0.00	0.00	25.00	25.00	0.00	0.00	25.00	6.17	25.00	0.00
0.00	0.00	25.00	25.00	25.00	0.00	25.00	0.00	0.00	0.00
0.00	0.00	13.41	25.00	0.00	0.00	0.00	0.00	0.00	0.00

$\mathbf{K}_{\mathrm{HL}}$					$\mathbf{K}_{\mathrm{MSD}}$				
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	9.44	25.00	25.00	25.00	25.00
0.00	25.00	25.00	0.00	0.00	0.00	25.00	25.00	25.00	25.00
22.03	25.00	25.00	0.00	0.00	0.00	0.00	0.00	25.00	0.00
0.00	25.00	25.00	0.00	0.00	0.00	0.00	0.00	9.39	0.00

References

1. Duda RO, Hart PE, Stork DG (2001) *Pattern Classification.* J. Wiley & Sons, New York
2. Konar A (2005) *Computational Intelligence (Principles, Techniques and Applications).* Springer, Berlin Heidelberg New York
3. Mangasarian O L (1965) *Operations Research* 13:444–452
4. Guarracino MR, Cifarelli C, Seref O, Pardalos P (2007) *Optimization Methods and Software* 22(1):73–81
5. Kiendl H, Knicker R, Niewels F (1996) Two-way fuzzy controllers based on hyperinference and inference filter. *Proc. 2nd World Automation Congress, vol. 4, Intelligent Automation and Control*, 387–394. TSI Press, Montpellier
6. Krone A, Schwane U (1996) Generating fuzzy rules from contradictory data of different control strategies and control performances. *Proc. 5th IEEE Int. Conf. on Fuzzy Systems*, 492–497 New Orleans
7. Kosko B (1997) *Fuzzy Engineering.* Prentice Hall
8. Jasinevicius R (1988) *Parallel Space-Time Computing Structures* (in Russian) Mokslas, Vilnius
9. Jasinevicius R, Petrauskas V (2008) Fuzzy expert maps: the new approach. *Proc. IEEE Congress on Evolutionary Computation.* IEEE Press, Piscataway
10. Barauskas R, Krusinskiene R (2007) *Journal of Sound and Vibration* 308:612–624
11. Juodzbaliene V, Muckus K, Krisciukaitis A (2002) *Acta Kinesiologiae Universitatis Tartuensis* 7:89–93
12. Grabauskas V (1991) *Medicinos enciklopedija, vol. 1.* Enciklopedij Leidykla, Vilnius

Table 6 Histograms' classification according to the T-S reasoning, when HL class is divided in to three subclasses for the PWLPP.

		Similarity function						Decision results
								Decision rules
		$\Phi_{\mathrm{HL}_1} =$ 92958.69	$\Phi_{\mathrm{HL}_2} =$ 79602.76	$\Phi_{\mathrm{HL}_3} =$ 78429.23	$\Phi_{\mathrm{MSD}_1} =$ 94535.19	$\Phi_{\mathrm{MSD}_2} =$ 76054.38	$\Phi_{\mathrm{MSD}_3} =$ 78433.33	
	c_i	1	1.17	1.19	0.98	1.22	1.19	
	Objects No.	$c_1\Phi_{\mathrm{HL}_1}$	$c_2\Phi_{\mathrm{HL}_2}$	$c_3\Phi_{\mathrm{HL}_3}$	$c_4\Phi_{\mathrm{MSD}_1}$	$c_5\Phi_{\mathrm{MSD}_2}$	$c_6\Phi_{\mathrm{MSD}_3}$	$\max(c_1\Phi_{\mathrm{HL}_1}, c_2\Phi_{\mathrm{HL}_2}, c_3\Phi_{\mathrm{HL}_3}) = A$ $\max(c_4\Phi_{\mathrm{MSD}_1}, c_5\Phi_{\mathrm{MSD}_2}, c_6\Phi_{\mathrm{MSD}_3}) = B$ $\max(A, B)$
HL	1	50771.76	9447.38	74766.51	53356.97	78604.91	40569.18	HL
class	2	33858.82	-8863.27	49209.19	60580.57	81035.74	54435.95	MSD
	3	74366.95	25050.00	33870.67	27887.61	42766.18	18109.76	SV
	4	92958.69	58096.11	54868.26	18592.65	57004.42	16361.60	HL
	5	-3542.02	-35839.04	31919.00	84667.09	109948.38	77102.78	MSD
	6	63563.34	52051.98	-26594.99	84000.61	89832.11	98595.59	MSD
	7	36497.10	6868.91	76255.58	44040.89	56745.98	27887.61	HL
	8	43192.09	16328.76	85352.42	53618.28	70534.90	42954.71	HL
	9	41137.33	74366.95	-24965.26	21114.87	9557.41	68915.32	HL
	10	70216.54	22266.02	87676.23	1013.55	40153.35	-20845.98	HL
	11	66349.70	92777.52	5910.65	13002.64	3384.19	60177.77	HL
	12	84708.15	125837.48	3391.98	-15828.78	-31880.94	53391.83	HL
	13	60346.70	92958.69	37438.45	55661.40	27887.61	73417.17	HL
	14	56369.02	37238.05	92958.69	26356.72	45774.13	11484.53	HL
	15	63916.98	5676.16	116515.67	14393.13	67796.39	-8083.02	HL
MSD	16	3193.04	-13729.77	-9023.57	93799.25	102319.06	91462.75	MSD
class	17	31543.31	28382.15	45972.29	65229.87	68936.07	69748.20	MSD
	18	67171.95	36563.95	-9764.36	66880.22	73193.83	80055.28	HL
	19	36434.47	60001.60	12962.92	60434.17	47635.27	70856.82	HL
	20	27887.61	27887.61	27887.61	92611.43	84165.07	89722.46	MSD
	21	76066.25	77104.70	31924.52	63282.16	51676.62	63010.10	HL
	22	39344.97	-23168.60	16016.06	58235.89	81626.26	58643.38	MSD
	23	29221.57	-29007.52	-11148.99	82361.38	95691.47	80125.02	MSD
	24	59034.40	32914.14	6918.12	82809.52	89127.24	82554.67	MSD
	25	35369.55	1360.60	60588.69	50898.48	71068.24	38615.59	HL
	26	10751.76	6617.15	39447.38	77674.38	84228.63	72717.59	MSD
	27	19295.18	20496.45	62703.50	55910.15	51502.06	51265.55	HL
	28	52715.84	57167.46	-7246.69	61760.93	63954.59	78643.79	MSD
	29	60513.27	71101.34	73655.56	55199.24	39875.27	46080.33	HL
	30	27231.68	24085.99	24934.00	92958.69	92958.69	92958.69	MSD
Number of errors								9

Dynamic Data-Driven Fuzzy Modeling of Software Reliability Growth

Olga Georgieva[1]

Abstract The paper deals with a new model description of software reliability growth dynamics. The model is applicable to multi-stage reliability growth that covers the complex defect detection rate and is based on Takagi-Sugeno fuzzy inference engine. The identification procedure determines model structure in real time based on evolving clustering algorithm. The clusters are discovered according to Gustafson-Kessel distance metric that copes with clusters of different shape and orientation. The developed model is validated through a case-study data set.

1 Introduction

Continuous availability of designed functionalities is a crucial qualitative characteristic of the software product. This quality is expressed by software reliability. The complexity of the contemporary software systems often does not guarantee full reliability at least for the reasonable time period of development and testing. Usually, the product is shipped if an acceptable low number of software failures could be discovered during the system operation. The effective assessment of the failure number is provided by a software reliability growth model. The model is built based on defect detecting data collected during the software testing process and enables to predict the number of defects remaining in the software. The goal is to reach an acceptably low defect discovery rate, which will guarantee software suitable for delivery.

The standard solution of this task statistically interpolates defect detection data to a mathematical function. As the phenomenon presents exponential character, different variations of the exponential function have been intensively explored [14], [6]. In order to account certain uncertainties this

[1] Department of Software Engineering, Faculty of Mathematics and Informatics, Sofia University, Bulgaria, o.georgieva@fmi.uni-sofia.bg

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 241–252.
springerlink.com

approach supposes normal distribution of the model parameters' values. However, statistics does not cover inexplicit information due to the subjectivity of the human behavior, which has large impact on the processes of software development, testing and system operation. We should take into account that software failures are result of designers, developers and/or testers' mistakes.

Seeking effective modeling techniques investigations have been directed to approaches able to grasp the uncertainty of fuzzy type [5], [10], [15]. Recently, fuzzy logic approach was explored as a powerful framework for software reliability growth description [12]. In order to account for the existing deterministic information Takagi-Sugeno (TS) fuzzy rule base scheme [3], [13] was successfully applied. This modeling approach tries to decompose the input-output data space into subspaces having vague boundaries and to approximate the system behavior in every subspace by a linear model. The proposed fuzzy model of software growth consists of a collection of linear submodels that represent the expected software faults as a function of historical measured data [1], [8].

Improvement of the reliability growth model could be search in a solution that accepts fuzzy clustering as an efficient approach for recognition of the TS model structure. Examples of clustering algorithms that has been explored in TS model identification procedure are Mountain clustering algorithm and its modification - Subtractive clustering algorithm as well as objective function clustering [3].

On the other hand the reliability has to be predicted using the data collected till the current moment. Thus, the proper task is to model software reliability in real time. In order to solve this problem on-line version of TS modeling algorithm should be investigated. Some authors solve this task by applying on-line extension of the Mountain/Subtractive clustering. They utilize recursive and noniterative technique for calculating the potential of the new data point in order to update the existing clusters or to discover new ones [2]. However, this solution reduces the model accuracy as it does not consider the clusters shape.

In this paper we propose new modeling approach for software reliability growth description based on TS model. The identification algorithm is realized through evolving clustering procedure [7] that copes the advantages of the objective function clustering enabling to identify clusters with a generic shape and orientation. It uses Gustafson-Kessel (GK) distance measure [9] to find elliptic clusters with different shape and orientation adapted to cover the individual character of the clustered data.

The TS fuzzy rule base system of software reliability growth model is introduced in the second section. Identification procedure is revealed in the third section. Real time identification algorithm is described in detail in the fourth section. The model efficiency is evaluated using a case study data set (fifth section). Concluding remarks are given in the sixth section.

2 Takagi-Sugeno Fuzzy Rule Base of Software Reliability Growth

General description of TS fuzzy model consists in a rule base that has a fuzzy antecedent part and a functional-type consequent of the following form:

$$\begin{aligned} &\text{If } \mathbf{x_k} \text{ belongs to cluster } Cl_i \\ &\text{then } y_{ik} = a_{i1}x_{k1} + a_{i2}x_{k2} + ... + a_{in}x_{kn} + b_i, \end{aligned} \tag{1}$$

where $\mathbf{x_k} = [x_{k1}, x_{k2}, \ldots, x_{kn}]$ is the value of the input vector at the current time instant k, y_{ik} is the output of the i-th linear subsystem function that has coefficient vector $Q_i = [a_{i1}, a_{i2}, \ldots, a_{in}, b_i]$ and c is the number of rules of the model rule base. By Cl_i, $i = 1, \ldots, c$ we denote the i-th cluster defined in the data space. Every cluster defines a rule and every rule corresponds to a certain subsystem.

The global process output y_{ik} is obtained by the following sum:

$$y_k = \sum_{i=1}^{c} \beta_i(\mathbf{x_k}) y_{ik}. \tag{2}$$

The parameter $\beta_i(\mathbf{x_k})$ is estimated as

$$\beta_i(\mathbf{x_k}) = \frac{w_i(\mathbf{x_k})}{\sum_{i=1}^{c} w_i(\mathbf{x_k})}, \tag{3}$$

under constraints

$$\sum_{i=1}^{c} \beta_i(\mathbf{x_k}) = 1 \text{ and } \beta_i(\mathbf{x_k}) \geq 0, \tag{4}$$

where $w_i(\mathbf{x_k})$ is the degree of fulfilment (DOF) of the i-th rule to the whole rule system output calculated for the input vector $\mathbf{x_k}$. The estimation of $w_i(\mathbf{x_k})$ is accomplished by multidimensional antecedent membership function method [3]. DOF is computed directly for the entire antecedent vector without decomposition. The distance between the antecedent vector and cluster centers is defined by recalling the cluster covariance matrix that includes all but not the last column of the cluster covariance matrix i.e. using only the antecedent covariance matrix.

The description given by the equations (1)-(4) provides gradual and smooth switching between distinct subsystems, which suits well to the real transition between different regions of the whole system behaviour.

2.1 TS Model Description of Software Reliability Growth

Software reliability growth has been grouped into two classes of models — concave (exponential type) that are continually bending downward and S-shaped. Both types of models have the same asymptotic behavior as the total number of defects detected asymptotically approaches a finite value [14]. Theoretical considerations given in the present section are provided for both model types. However, further on practical investigations are focused on the concave dynamics. The concave model type is of real practical interest as most data sets accumulated during the test process exhibit this behavior.

Concave behavior of the reliability growth expresses exponential growth which passes through two phases. In the beginning the process operation detects failures relatively frequently, which forms the phase of fast growth. In the second process phase the rate of failure detection decreases and the dynamics appears to be in (or close to) the steady-state. Within each phase the reliability growth could be linearly approximated. According to this preliminary information we construct dynamic TS fuzzy if-then rule base in the following regression form:

$$\begin{aligned} &\text{If } \mathbf{x_k} = [x_k, x_{k-1}] \text{ belongs to cluster } Cl_i \\ &\quad \text{then } x_{ik+1} = a_{i1}x_k + a_{i2}x_{k-1} + b_i, \end{aligned} \tag{5}$$

where x_k is the number of failures in the k-th moment, $i = 1, 2$ and cluster vector is $\mathbf{Cl}$=[Cl_1, Cl_2]=[*Exponential*, *SteadyState*].

In case of S-shaped growth the model has three rules that correspond to the respective three phases of the reliability growth dynamics namely lag, exponential and steady-state phases. In this case cluster term set is $\mathbf{Cl}$=[Cl_1, Cl_2, Cl_3]=[*Lag*, *Exponential*, *SteadyState*].

The global model output predicts the failure number:

$$x_{k+1} = \frac{\sum_{i=1}^{c} w_i(\mathbf{x_k}) x_{ik+1}}{\sum_{i=1}^{c} w_i(\mathbf{x_k})}, \tag{6}$$

where $c = 2$ for the concave model and $c = 3$ for the S-shaped model, $w_i(\mathbf{x_k})$, $i = 1, c$ is the DOF of the i-th rule output.

The model predicts the failure amount that would be accumulated in the next time instant. The obtained value is a non integer one. In order to get a usable result we suggest to assume the closest integer value.

2.2 Multi-stage Model Description of Software Reliability Growth

The model described by equations (5)-(6) assumes that the defect detection rate decreases as the testing process prolongs. In fact if each defect is fixed as it has been discovered the defect amount in the code should decrease and revealed dependence will follow asymptotical dynamic of concave or S-shape form. However, more realistic consideration of the testing process should take into account the repairing effects. For instance, if a significant amount of new code is added in order to fix the failure, the number of rest defects could increase. In this case we can accept a new model of the defect detection rate. We also could suggest that the defect detection rate has the same nature and the new model has the same structure as the current one. Main obstacle in this assumption is how to define appropriate model switching. Real practical concern is to decide whether new dynamics (model) is appeared and to identify the coefficients of the new model.

The idea we are realizing here is to identify the repairing phenomenon online by increasing the number of rules of the basic growth description model (5) in real time mode. It is accomplished by on-line identification procedure that automatically detects the new phases in the software reliability growth occurred if a certain amount of new code is added.

3 Model Identification

Identification of the model (5)-(6) needs interdependent procedures of structure and parameter identification that are revealed in the next subsections 3.1 and 3.2, respectivly.

3.1 Structure Identification

The structure identification of the model aims to determine the number and types of the fuzzy subsystems. A widely applied method is based on fuzzy clustering procedure [3].

Let raw data are provided in a matrix form $Z_{N\times(n+1)} = [\mathbf{z_k}], k = 1, \ldots, N$ with $\mathbf{z_k} = [x_{k1}, x_{k2}, \ldots, x_{kn}, y_k] = [\mathbf{x_k}, y_k]$ being $n+1$ dimensional vector and N — the number of data points. The dimensionality of the introduced basic model of software reliability growth is $n = 3$, where the output vector equals to the predicted number of failures $y_k = x_{k+1}$.

Fuzzy clustering is based on iterative optimization of the objective functional [4]:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^{m} d_{ik}^{2}. \tag{7}$$

Here μ_{ik} is the membership degree of k-th data to the i-th cluster. The clustering obtains a vector $\mathbf{v} = [v_1, v_2, \ldots, v_c]$ of cluster centers. The scalar parameter m determines the fuzziness of the resulting clusters. Usually $m = 2$ [3], [4].

Oblong clusters with different orientation in the space typically characterize the phases of the reliability growth. Most appropriate distance norm in this case is Gustafson–Kessel [9]:

$$d_{ik}^{2} = (\mathbf{z_k} - v_i) A_i (\mathbf{z_k} - v_i)^T, \tag{8}$$

where the norm inducing matrix A_i is symmetric and positive:

$$A_i = [\det(F_i)]^{1/n} F_i^{-1}. \tag{9}$$

Main feature of the obtained partition is the local adaptation of the distance metric to the shape of the cluster according to the covariance matrix F_i:

$$F_i = \frac{\sum_{k=1}^{N} \mu_{ik}^{m} (\mathbf{z_k} - v_i)^T (\mathbf{z_k} - v_i)}{\sum_{k=1}^{N} \mu_{ik}^{m}}. \tag{10}$$

Batch application of the clustering procedure i.e. when data are off-line available, alternatively iterates to determine the membership degree

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}^2}{d_{jk}^2})^{\frac{1}{m-1}}} \tag{11}$$

and the cluster center

$$v_i = \frac{\sum_{k=1}^{N} \mu_{ik}^{m} \mathbf{z_k}}{\sum_{k=1}^{N} \mu_{ik}^{m}}. \tag{12}$$

3.2 Parameter Estimation

As model structure is recognized the model coefficients have to be estimated interdependently on the clustering parameters (11) and (12). The vector of consequent coefficients $Q = [Q_1, \ldots, Q_c]$ appears linearly in the matrix model output description. Its value could be easily defined by applying the linear least square method (LLSM). For instance, the authors of [3] accepted LLSM to solve the off-line modeling task. Recursive version of LLSM is accepted in case of adaptive model implementation [2].

4 Real Time Model Identification

The structure identification of software reliability growth description needs clustering that deals with streams of data and recognizes the system structure in real time. Evolving clustering intends to deal with this task. Let us underline that the evolving clustering algorithm works with the currently available data $\mathbf{x_k}$.

The introduced evolving clustering algorithm [7] assumes that the boundary of each cluster is defined by a cluster radius. The radius r_i of the i-th cluster is equal to the maximal distance between the cluster centre v_i and the points belonging to this cluster with a membership degree larger or equal to a given threshold membership degree μ_h:

$$r_i = max \left\| v_i - \mathbf{x_j} \right\|_{A_i} \text{ for } \forall x_j \in i^{th} \text{ cluster and } \mu_{ij} \geq \mu_h, \tag{13}$$

where $\|.\|_{A_i}$ is the GK distance norm determined according to equation (8) for which the data $\mathbf{x_j}$ belongs to the i-th cluster with membership degree μ_{ij} such that $\mu_{ij} \geq \mu_h$.

Three possibilities should be evaluated if a new data $\mathbf{x_k}$ has been currently measured. First, the data belongs to an existing cluster if it is within the cluster boundary. This case imposes just clusters' update. If the data point is not within the boundary of any existing cluster a new cluster is a subject of assessment. Alternatively, $\mathbf{x_k}$ is an outlier, which does not affect neither the data structure neither clusters' parameters. In order to assess a new cluster we serve the following considerations.

The minimal distance d_{pk} determines cluster p closest to the current data:

$$p = arg\ min_{i=1,\ldots,c}(d_{ik}). \tag{14}$$

The data $\mathbf{x_k}$ is assigned to the cluster p if the distance d_{pk} is less or equal to the radius r_p

$$d_{pk} \leq r_p. \tag{15}$$

The Kohonen rule [11] to update the p-th cluster parameters' values is applicable:

$$v_{p,new} = v_{p,old} + \alpha(\mathbf{x_k} - v_{p,old}) \tag{16}$$

$$\mathbf{F_{p,new}} = \mathbf{F_{p,old}} + \alpha((\mathbf{x_k} - v_{p,old})^T(\mathbf{x_k} - v_{p,old}) - \mathbf{F_{p,old}}). \tag{17}$$

If condition (15) fails a new potential cluster is assessed. For this the number of clusters is incremented

$$c = c + 1 \tag{18}$$

and the incoming data $\mathbf{x_k}$ is accepted as a center of the new cluster v_{new} with a covariance matrix $\mathbf{F_{new}}$ initialized by the covariance matrix of the closest cluster:

$$v_{new} = \mathbf{x_k}, \mathbf{F_{new}} = \mathbf{F_p}. \tag{19}$$

In order to quantify the credibility of the estimated cluster a parameter P_i is introduced to assess the number of points belonging to the i-th cluster. Its lower bound could be estimated from the minimal number of data points necessary to learn the parameters of the covariance matrix as:

$$P_{min} = n(n+1)/2. \tag{20}$$

Alternatively, it could be context defined but the value should not be less than P_{min}.

Appropriate choice of the membership threshold μ_h, coefficient α and credibility parameter P_{min} depends on the density of the data set and level of cluster overlapping. Their value is of a real concern for the model identification. We assume default value of the membership threshold μ_h=0.5 that balances between having crisp cluster separation and most tolerant fuzzy cluster separation. For a stricter identification of proper clusters the predefined threshold membership degree should be chosen larger. For more tolerant identification it should be chosen smaller. The learning rate α determines the step of searching. A large value guarantees sparsely selected clusters, which could ignore a valuable cluster and the recognized rules may not cover accurately the data space.

5 Benchmark Model Development

The developed software reliability growth model was evaluated on a bench mark data set provided by a case study investigation of TANDEM Computers company [14]. The data base consists of data sets of four separate software releases in which for every test week the number of failures was counted.

Release 2 and Release 3 present certain overlapping. Release 2 was a preliminary release used by very few customers. It was tested for nineteen weeks. Release 3 was very similar to Release 2 with some functionality and performance enhancements due to a new functionality being added in the test week 17 of Release 2. Thus, in order to examine predicting abilities of the model, Release 2 was treated separately. Then data of both releases were merged suitably in one release data set named Release 23. This data set forms a multi-stage reliability growth dynamics that passes through consequently two concave processes.

In order to initialize the identification algorithm first cluster was defined off-line over data of the fast growth phase of each release. Default value $\mu_h = 0.5$ was accepted. The learning rate was fixed at $\alpha = 0.05$ and minimal cluster credibility was set to $P_{min} = 5$. In order to increase the algorithm sensitivity in some of the simulations the parameter value was decreased to the minimal acceptable value $P_{min} = 3$.

Simulation of Single Release Models

TS type models in the form of the rule system (5)-(6) have been identified separately for respective release data sets. Simulation curves present concave behavior (Fig. 1). Due to the specificity of the task the overestimation of the data is preferred than the underestimation. It can be noticed that the model accuracy is acceptable in the beginning of the process when high detecting rate occurs. In the phase of asymptotical behavior of each model the process is better fitted than in the first process stage. It is explained by the increased number of data, that benefit the accuracy of the identified model. The increased predictability in the second phase is of a large benefit as we are interested in the number of remaining failures after certain test time.

The model performance was evaluated through model performance indexes as detected maximal absolute error (Max error) and root mean square error (RMSE) presented in Table 1. Cluster centers values defined off-line as well as their final values at the real time simulation are both presented.

Table 1 Performance indexes and obtained cluster centers

Indexes	Release 1	Release 2	Release 4	Release 23
Max error	3.5846	4.0482	3.2948	4.0727
RMSE	1.51	1.72	1.32	1.92
$v_{1\,off\,line}$	[33.80 39.91 46.19]	[28.62 36.31 45.28]	[16.77 18.40 22.40]	[28.75 36.48 45.44]
$v_{2\,off\,line}$	[86.99 90.53 93.21]	[98.24 102.95 106.8]	[28.06 31.13 32.32]	[98.36 103.0 106.9]
$v_{2\,on\,line}$	[93.99 96.61 98.33]	[102.7 106.4 111.5]	[30.90 33.66 33.84]	[98.76 104.9 110.4]
$v_{3\,on\,line}$	-	-	-	[174.4 177.3 179.2]

Simulation of Multi-stage Model

Identification of TS model of the merged release Release 23 was investigated as a case of multi-stage software reliability growth modeling. The real time algorithm recognizes the second cluster which copes the steady state phase of Release 2 (Fig. 2). By continuing the simulation, an additional cluster has been identified. It comprises data of Release 3. Thus, exponential and steady state phases of the second release were defined as well. Performance indexes (Table 1) show the model accuracy relevant to the single release models. Maximal error of Release 23 simulation is commensurable with the maximal error of Release 2 simulation. The value of variance account for index is VAF=99.84, which illustrates very high model fitting.

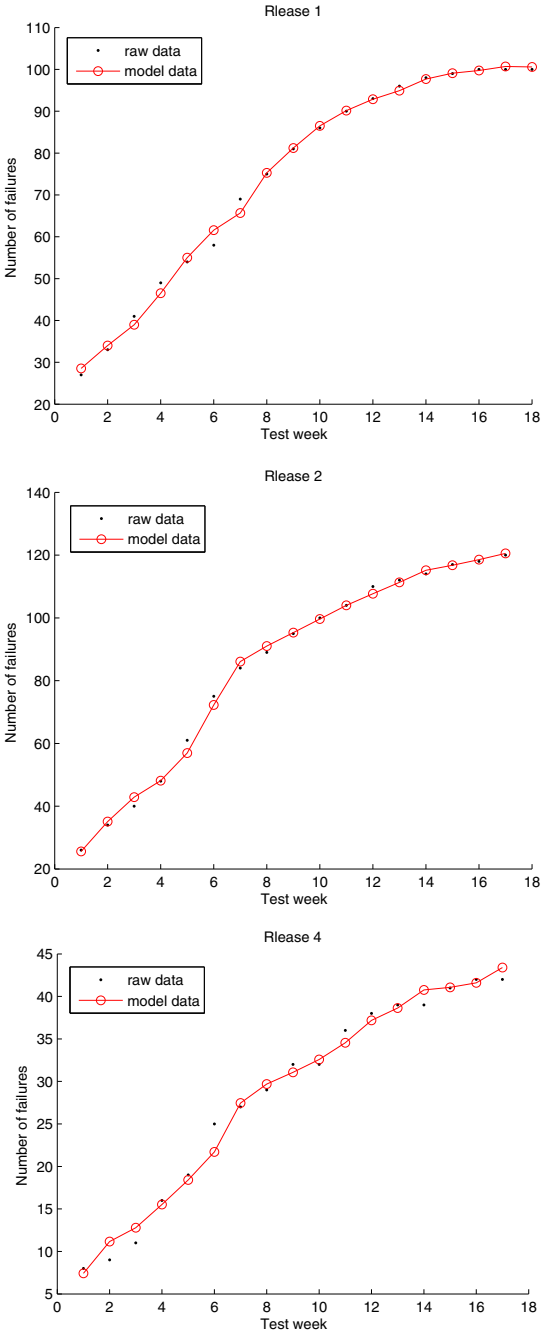


Fig. 1 Single stage model simulation

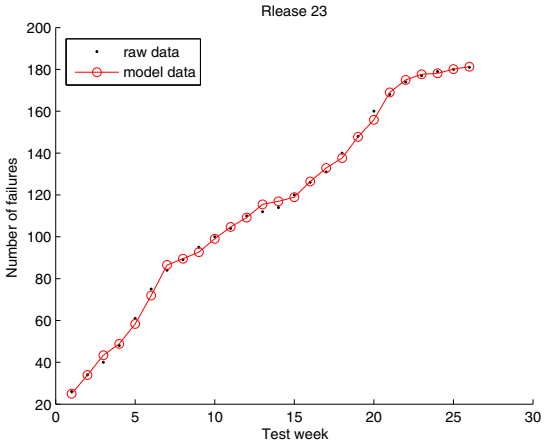


Fig. 2 Multi-stage model simulation

6 Conclusion

The paper introduces a new model description of the software reliability growth dynamics. It is based on fuzzy logic mathematical frame realized by Takagi–Sugeno fuzzy inference engine. The model combines both vague and deterministic information.

The model is able to cover complex growth dynamics result of repairing process when significant amount of new code is added during the test period. The proposed modeling scheme implements evolving clustering technique that recognizes the model structure in a real time mode. The model accuracy is reached for GK distance norm that was applied for cluster identification.

Acknowledgements This research has been partially supported by the Science Fund of Sofia University "St. Kl. Ohridski", Bulgaria under the project "Software Dependability Estimation via Data Mining Methods" 2011 − −2012 and the COST Action *IC* 0702.

References

1. Aljahdali S (2011) Development of Software Reliability Growth Models for Industrial Applications Using Fuzzy Logic. *Journal of Computer Science* 7(10):1574–1580
2. Angelov P, Filev D (2004) An approach to online identification of Takagi-Sugeno fuzzy models. *IEEE Trans. on SMC, Part B: Cybernetics* 34(1):484–498
3. Babuska R (1998) *Fuzzy modeling for control.* Kluwer Academic Publishers, Boston
4. Bezdek JC (1981) *Pattern recognition with fuzzy objective function algorithms.* Plenum Press, New York

5. Dimov A, Punnekkat S (2005) On the Estimation of Software Reliability of Component-Based Dependable Distributed Systems. In: *Quality of Software Architectures and Software Quality*, LNCS 3712:171–187
6. Farr W (1996) Software reliability modeling survey. In: Lyu MR (ed.) *Handbook of Software Reliability Engineering*, 71–117. McGraw-Hill, New York
7. Filev D, Georgieva O (2010) An Extended Version of Gustafson-Kessel Clustering Algorithm for Evolving Data Stream Clustering. In: Angelov P, Filev D, Kasabov N (eds.) *Evolving Intelligent Systems: Methodology and Applications*, 293–315. IEEE Press Series on Computational Intelligence
8. Georgieva O (2011) Takagi-Sugeno Type Fuzzy Logic Description of Software Reliability Growth. *Proc. Automatics and Informatics'11* B:117–120. Sofia, Bulgaria
9. Gustafson DE, Kessel WC (1979) Fuzzy clustering with a fuzzy covariance matrix. *Proc. IEEE CDC*, 761–766. San Diego
10. Junhong G, Xiaozong Y, Hsongwei L (2005) Software Reliability Nonliner Modeling and Its Fuzzy Evaluation. *Proc. 4th WSEAS International Conference*, 49–54. Sofia, Bulgaria
11. Kohonen T (1989) *Self-Organization and Associative Memory, 3rd edition*. Springer-Verlag, Berlin
12. Kumar R, Khatter K, Kalia A (2011) Measuring software reliability: a fuzzy model. *ACM SIGSOFT Software Engineering Notes* 36(6)
13. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modelling and control. *IEEE Trans. on System Man Cybernetic* 15(1):116-132
14. Wood A (1996) Software Reliability Growth Models. Tandem Technical Report 96.1., Part Number 130056, Tandem computers
15. Wu T H, Chen W (1995) Fuzzy Software Reliability Allocation. *Proc. of Computer Symposium*, 905–912

Dynamic Texture Recognition Based on Compression Artifacts

Dubravko Ćulibrk[1], Matei Mancas[2], and Vladimir Ćrnojevic[1]

Abstract The paper proposes a novel approach to the classification of compressed videos containing dynamic textures. The term dynamic texture is usually used with reference to image sequences of various natural processes that exhibit stochastic dynamics (e.g., water, fire and windblown vegetation). Description and recognition of dynamic textures have attracted growing attention.

Although one of the most important prospective applications of the technology is content-based video retrieval, recognition of dynamic textures for compressed video has not been considered. The content of video and dynamic textures in particular, profoundly affect the performance of video compression algorithms. The prominence of compression artifacts can, therefore, be used to recognize dynamic textures in compressed videos. In the paper, we show how features, previously proposed for quality assessment, statistical analysis and a soft computing technique (neural networks) can be used to discern 23 different classes of dynamic textures in a standard video database, with 99.5% accuracy.

1 Introduction

Dynamic textures represent a set of phenomena occurring in nature, where the perceived changes in the appearance of a system of large number elements are consistent, although the individual elements undergo stochastic changes in theirs. Typically the changes are due to motion (e.g. turbulent water water, smoke, vegetation in the wind, insect swarms), but may be the result of the changing intensity of light emitted (e.g. fire). In the computer vision literature, such patterns have appeared collectively under various names,

[1] University of Novi Sad, Serbia, {dculibrk,crnojevic}@uns.ac.rs
[2] University of Mons, Belgium, matei.mancas@umons.ac.be

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 253–266.
springerlink.com

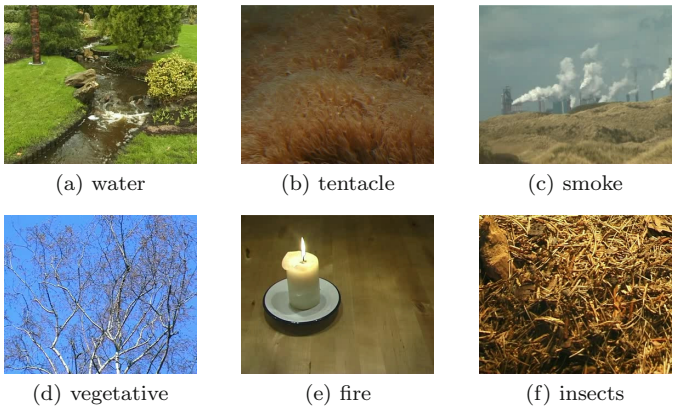
including, turbulent flow/motion, temporal textures, time-varying textures, dynamic textures, and textured motion [6]; the term dynamic texture will be used herein. Zhao and Pietikinen consider such phenomena extensions of the static texture to the temporal domain [26], since the effect is that of a textured object undergoing transformations. Derpanis and Wildes [6], however, point out that the term can apply equally well to simpler phenomena when analyzed in terms of aggregate regional properties (e.g., orderly pedestrian crowds and vehicular traffic).

The ability to recognize dynamic textures based on visual processing is of significance to a number of applications, including, video indexing/retrieval, surveillance and environmental monitoring where they can serve as keys, isolate background clutter (e.g., fluttering vegetation) from activities of interest and detect various critical conditions (e.g., fires), respectively. It comes as no surprise that a significant amount of research effort has been directed toward solving this problem [4] [26] [12] [6]. However, to the best of our knowledge, no one has dealt with the possibility of recognizing dynamic textures in compressed (coded) videos, although this is the 'natural' state of the material in applications such as content-based video retrieval and the preferred way of storing and transmitting visual data in all other.

The quality of coded video sequences depends on the video codec, bit-rates required and the content of video material [5]. Clearly, if the bit-rate and the codec are the same over a range of sequences - a reasonable assumption for multimedia databases - the quality of compressed videos is dependent only on the content. We propose a video classification approach that exploits this relationship. Using the compressed videos available in a standard database used for dynamic texture recognition [12], we show that the measures of the level of artifacts introduced by the coding algorithm can be used as basis for efficient dynamic texture recognition.

Based on video quality measures, content-dependent features are extracted for the frames of the video. These are then aggregated so that each video sequence is represented by a fixed-length signature derived from the feature-values obtained for single frames. Video Signatures (VS) are subsequently used to train a boosted soft computing (neural network) classifier [2]. Experimental results obtained through cross-validation show that the classifier is able to achieve perfect (99.5% accurate) classification for the data set used.

The paper makes several contributions. Dynamic texture classification from compressed video is considered for the first time. To the best of our knowledge, no one has attempted to use the correlation between coding artifacts, video quality and content to classify dynamic textures. The proposed methodology relies on a state-of-the-art Video Quality Assessment (VQA) approach and exploits visual saliency due to motion to extract dynamic-texture related changes. Finally we propose the use of boosting Multi Layer Perceptron (MLP) neural networks to classify dynamic textures - another novelty- and show that such a classifier, combined with the proposed features, can



(a) water (b) tentacle (c) smoke

(d) vegetative (e) fire (f) insects

Fig. 1 Sample frames from DynTex sequences

achieve 99.5% accurate classification for 23 classes of dynamic textures represented in a standard video data set.

The rest of the paper is organized as follows: Section 2 deals with the relevant published work. Section 3 describes the methodology proposed. Section 4 discusses experiments performed and results achieved. Section 5 is dedicated to our conclusions.

2 Related Work

2.1 Dynamic Texture Classification

The research into the classification and recognition of dynamic textures continues unabated [6] [26]. A large number of approaches have been proposed over the last ten years. In their 2005 survey Chetverikov and Péteri [4] divided the existing approaches into five classes: methods based on optic flow, methods computing geometric properties in the spatiotemporal domain, methods based on local spatiotemporal filtering, methods using global spatiotemporal transforms and model-based methods that use estimated model parameters as features. Regardless of the type of the approach, they attempt to extract features descriptive of the dynamic texture and classify them by either defining a suitable distance measure and creating a simple distance-based algorithm for comparison or training a machine learning algorithm to achieve the task.

In their 2007 paper [26], Zhao and Pietikinen proposed volume local binary patterns (VLBP) as features to describe dynamic textures. The VLBPs are an extension of the LBP operator widely used in ordinary texture analysis, that combine motion and appearance. They tested their approach using videos

generated by extracting parts of the sequences in the DynTex database [12], creating a data set that had 10 examples of a certain class derived from single DynTex sequences. Their classifier is a simple nearest neighbor classifier, based on the log-likelihood statistic that allows them to compare VLBPs, and they used leave-one-group-out (i.e. n/m fold cross-validation [22]) to measure performance, where m corresponds to the number of examples extracted for a single dynamic texture and n is the total number of examples. Various classification rates were achieved depending of whether or not the features used were shift-invariant and how long the feature vector was. Their best result is an accurate classification rate of 95.71%, achieved for a shift-invariant VLBP and a fairly large feature vector $(4,176$ bins) .

Chan and Vasconcelos [3] model the dynamic texture as a linear dynamic system (LDS) and achieve good classification using Martin distance to compare the models. They evaluated both nearest neighbor and support vector machine (SVM) classifiers and showed that the use of a machine learning algorithm such as SVM can improve the classification significantly. Through the use of the SVM classifier they achieved accurate classification rate of 97.5% on the UCLA database [15]. More recently (2009) their work has been extended by Ravichandran *et al.* [13] to use bags of LDSs to achieve improved view-invariant texture classification, when eight classes of textures are concerned.

Recently (2010), Derpanis and Wildes [6] proposed new features based on spatiotemporal oriented energy filters to describe dynamic textures and classify them. They identified 7 semantic categories in the UCLA database (flames, fountain, smoke, turbulence, waves, waterfall, vegetation) and achieved a comparatively low classification rate of 92.3%, on sequences derived from this database. However, they specifically considered shift-invariant recognition, and report improved performance under this conditions.

To the best of our knowledge no one has considered the problem of classifying dynamic texture in compressed videos nor the use of features used to measure different artifacts introduced by coding, as basis for dynamic texture classification/recognition.

2.2 Video Quality Assessment

The quality of coded video sequences depends on the video codec, bit-rates required and the content of video material [5]. If the bit-rate and the codec are the same over the range of sequences the quality of compressed videos is dependent only on the content, allowing for the use of features related to quality to discern content. This is not an unrealistic constraint, e.g. a quick calculation reveals that environ 290 million videos have been uploaded to YouTube in 2010 [24], all using the same codec and a significant subset with default parameter settings.

Overall degradation in the quality of the sequence, due to encoder/decoder implementations as part of transport stream at various bit rates, is a compound effect of different coding artifacts [21]. Three types of artifacts are typically considered pertinent to block coded data: blocking, ringing and blurring. Blocking appears in all block-based compression techniques, which include all contemporary codecs [14] [8], due to coarse quantization of frequency components [19] [20]. It can be observed as surface discontinuity (edge) at block boundaries. These edges are perceived as abnormal high frequency components in the spectrum. Blocking is usually masked by the presence of strong texture in the background and blockiness measures are designed to estimate what part of the discontinuity on the block edges is due to the blocking effect vs. the texture in the content [1]. In the setup proposed in this paper, blockiness measures are related to the texture in the background. Ringing is observed as periodic pseudo edges around original edges [11]. It is due to improper truncation of high frequency components. In the worst case, the edges can be shifted far away from the original edge locations, observed as false edge. Blurring, which appears as edge smoothness or texture blur, is due to the loss of high frequency components when compared with the original image. Blurring causes the received image to be smoother than the original one [7] and the measures of blurring try to estimate the difference in activity of the original content with respect to coded version. They are profoundly influenced by the textures in the video content.

A large number of published papers exist that propose different measures of prominent artifacts which appear in coded images and video sequences [19] [5]. In this study we limit ourselves to no-reference approaches, where only compressed video is available. This is a harder problem, but more realistic in applications such as video-retrieval.

Several published approaches to measuring video quality are of interest for the discussion in the following sections. Wang *et al.*[19] proposed a no-reference approach to quality assessment in JPEG coded images. Their final measure is derived as a non-linear combination of a blockiness, local activity and a so-called zero-crossing measure. The combination is supposed to provide information regarding both blockiness and blurring (via the two latter measures) in JPEG coded images. More recently, Babu *et al.* [1] proposed a blockiness measure for use in VQA, which takes effects along each edge of the block into account separately.

Measures related to various artifacts are usually evaluated for each frame of the sequence and collapsed temporally to arrive at a quality measure for the whole sequence [23] [18] [10] [5].

Recently, Culibrk *et al.* [5] proposed a VQA approach that improves quality estimation by separately considering the regions of the frame in which salient-motion is present and the rest of the frame. Using a simple multi-scale foreground-background segmentation approach, they detect the salient regions and calculate a number of features related to the observed temporal changes. In addition, they calculate the blockiness and blurring measures pro-

posed by Babu *et al.* [1] and Wang *et al.* [19] for the salient and non-salient parts of the frame. Using these features they train a neural-network and a decision tree classifier that are able to achieve state-of-the art quality estimation on a per-frame basis. The final estimate of the quality is the median value obtained for the frames of the sequence.

The approach of Culibrk *et al.* has been selected as a state-of-the art approach for measuring video quality that is used in the study presented here. Since dynamic textures are by their very nature salient due to motion, this approach enables us to capture the features related to the dynamic-texture part of the sequence frames and filter out the rest of the sequence.

2.3 Classification

Once descriptive features are extracted the preferred approaches to classification of dynamic textures seem to be the Nearest Neighbor (NN) classifier and Support Vector Machines (SVM) [26] [3] [13] [6].

Culibrk *et al.* [5], in the other hand, proposed using either a decision tree classifier or a Multi Layer Perceptron (MLP) [9] neural network to estimate the quality of video based on their features. In addition, they performed automatic feature selection to evaluate the impact of saliency and showed that an MLP estimator can achieve good results using a subset of just 5 features.

Here we propose using an MLP based classifier to discern different texture classes. Neural networks represent a class of machine learning algorithms designed to follow the basic principles of biological neural cells and as such fall into the domain of soft computing. They consist of a number of interconnected nodes that receive signals through their input connections, do simple processing and pass the output to other neurons. The connections between the neurons emulate the synapses between the neurons in biological systems and are assigned weights that code the relative influence between the connected nodes. In artificial neural networks the weights are learned from data in order to create classifiers or estimators with the desired behavior.

To enhance the performance of the MLP classifier we propose using a meta-learning algorithm Adaptive Boosting [16]. The effect of such an approach is analogous to creating a cascade of neural network classifiers, each trained on the set of examples that are incorrectly classified by the preceding stages. Schwenk and Bengio [17] discuss the merits of such an approach in detail.

3 Proposed Approach

A block diagram of the proposed dynamic texture classification approach is shown in Fig 2. The input data are compressed videos of dynamic textures. If the video is not compressed it can easily be coded with any lossy compression algorithm. Each video in the data set is initially processed to extract measures related to motion, salient changes, blurring and blockiness. A total of 17 measures is extracted for half of the frames of video, distributed uniformly - once the measures have been calculated for a frame, the next frame is skipped. This increases the efficiency and has no impact on the effectiveness. The values of measures for all frames of a single video are clustered into 10 clusters using k-means clustering [25]. The process yields 10 cluster centroids that represent each video. The set of centroids is a fixed-size representation of a video, regardless of the number of frames it has. This is referred to as a *Video Signature* (VS).

Once the signatures for all the videos are computed they are used to train and test a neural-network-based meta classifier. Each centroid is used as a separate input the classifier and each video is represented by 10 centroids comprising the signature. This is done to make the approach less sensitive to the measurement error introduced by the saliency-detection module, which takes approximately 50 frames to adapt to sudden scene changes and to learn the background model when presented with a new sequence.

The classifier is used to discern the semantic class for each of the 10 centroids comprising a VS. The mode (most common value) of the 10 class labels obtained in this manner is used the final classification of the video.

3.1 Extracting Features

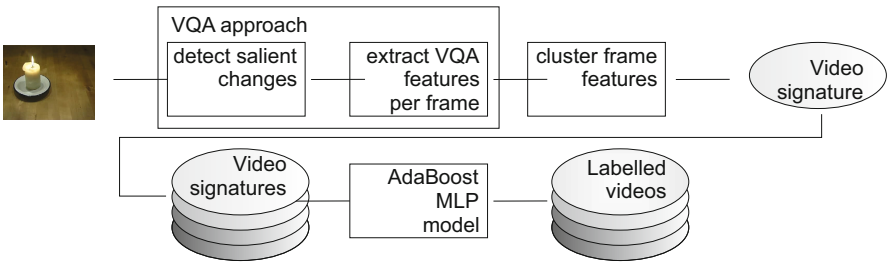


Fig. 2 Proposed video description approach: A video signature is extracted for each video, the signatures used to train a boosted MLP classifier that can be used to classify other videos

3.2 Basic Features for Video Classification

Set of features related to video quality was adopted from the work of Culibrk *et al.* [5].

The approach proposed in [5] attempts to estimate salient motion in each frame of the sequence by performing background subtraction at several different scales. The scaled frames are obtained from a frame of the sequence by performing spatial Gaussian filtering and decimating the frame to get the next scale. This yields a representation of each frame of the sequence in the form of a Gaussian pyramid. The same process is applied to background frames. The results of background subtraction at each scale are thresholded to eliminate small changes and summed up to form a single saliency map. Outlier detection is then used to determine which parts of the map are salient and which are not. Even with a small number of scales (3-5), the approach is able the achieve meaningful, if somewhat coarse, segmentation of interesting moving objects in the scene. In the case of the DynTex database, this corresponds to the dynamic-texture regions of the frame. The process is illustrated in Figure 3.

Once the salient parts of the frame have been determined several basic features are used to describe the salient motion in a frame: number of salient regions, their average size, and first moments (mean and standard deviation) of the difference between the current frame and background frames, calculated separately for salient and non-salient regions.

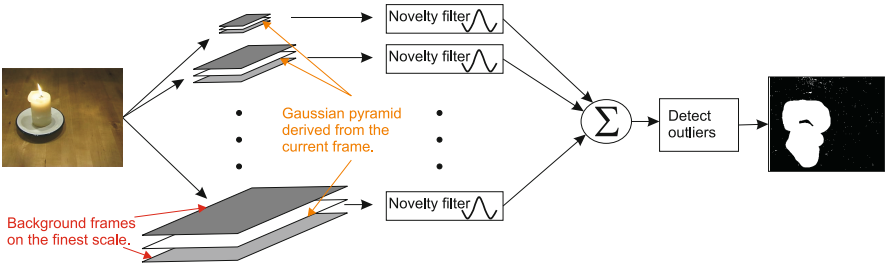


Fig. 3 Salient-motion region segmentation.

Also, to account for blurring and blockiness, Z-score measures proposed by Wang *et al.* [19] and the blockiness measure proposed by Babu *et al.* [1] are calculated separately for salient, non-salient and (in the case of the last feature) border regions. This should provide a good description of the texture within the different regions of the frame.

The blockiness measures proposed by Wang *et al.* and Babu *et al.* are profoundly different. Babu *et al.* focus on the effects that can be observed along the edges of a single block. Their measure is designed to detect blocks

with low spatial activity along the edges, but significant differences across them.

To characterize the activity on the inside of the block edge they calculate the standard deviation of pixel values for 6-pixel long stretches along he border of the block, since they observed that blockiness that spans less than 6 pixels is not perceived as significant. For each edge of the block they try to detect if there is significant activity that could mask the blockiness effect. Let $\{I_{k,j}|k \in [1,4], j \in [1,8]\}$ be the edges of a block and $\{O_{k,j}, k \in [1,4], j \in [1,8]\}$ the corresponding pixels across the edge of the block. We first consider the standard deviation of pixel values on the inside of block edges:

$$\sigma_{k,j} = stddev(I_{k,j}), k \in [1,3], j \in [k, k+5] \tag{1}$$

Then we compute the gradient across block edges for each subsegment of the edge:

$$\Delta_{k,j} = mean(|I_{k,j} - O_{k,j}|), k \in [1,3], j \in [k, k+5] \tag{2}$$

If any of $\sigma_{k,i}$ is below an empirically selected threshold ε, than that edge can contribute to the blockiness, but it will do so only if the gradient is larger than a different threshold τ. For a block i of a frame, we define

$$W_i = \begin{cases} 1, (\exists \sigma_{k,j}, \Delta_{k,j})(\sigma_{k,j} < \varepsilon \wedge \Delta_{k,j} > \tau) \\ 0, \text{otherwise.} \end{cases} \tag{3}$$

Finally, we calculate the proportion of blocks that contributes to the blockiness effect as the measure of blockiness:

$$BB = \frac{\sum_{i=1}^{NB} W_i}{\text{NB}} \tag{4}$$

where NB is the number of blocks in the region considered.

The authors of the approach [1] suggest $\varepsilon = 0.1$ and $\tau = 2.0$, which are also the values used in the study presented here.

The approach of Wang *et al.* is based on the observation that the artifacts can be detected if the image is transformed to the frequency domain and its power spectrum examined. They design their measures of blurring and blockiness in an attempt to achieve a less computationally intensive approach than that of computing the full power spectrum. Let $x(m,n)$ $m \in [1,M]$ and $n \in [1,N]$, be the pixel values (signal) for a frame. First a differencing signal is calculated along the horizontal lines:

$$d_h(m,n) = x(m,n+1) - x(m,n), n \in [1,N-1] \tag{5}$$

The blockiness measure proposed by Wang *et al.* tries to take into account the differences between a whole line of blocks, rather than looking at a single block:

$$B_h = \frac{1}{M(\lfloor N/8 \rfloor - 1)} \sum_{i=1}^{M} \sum_{j=1}^{\lfloor N/8 \rfloor - 1} d_h(i, 8j) \tag{6}$$

Thus, the Wang *et al.* provides a more wider-range measure of blockiness, when compared to the basic Babu *et al.* metric.

Wang *et al.* proposed two measures in an attempt to characterize the spatial activity of the signal. Their motivation lies in the fact that activity is reduced by blurring. The activity is related to how pronounced the texture is in a particular region of the frame. The first measure is the average absolute difference between in-block image samples:

$$A_h = \frac{1}{7} \left[\frac{8}{M(N-1)} \sum_{i=1}^{M} \sum_{j=1}^{N-1} |d_h(i,j) - B_h| \right] \tag{7}$$

The second measure is the zero-crossing (ZC) rate. They define for $n \in [1, N-2]$:

$$z_h(m,n) = \begin{cases} 1, \text{ horizontal ZC at } d_h(m,n) \\ 0, \text{ otherwise} \end{cases} \tag{8}$$

the horizontal ZC rate can then be estimated as:

$$Z_h = \frac{1}{M(N-2)} \sum_{i=1}^{M} \sum_{j=1}^{N-2} z_h(m,n) \tag{9}$$

The vertical features (B_v,A_v and Z_v) are then calculated in a similar fashion. The overall blockiness, activity and ZC rate are calculated as:

$$B = \frac{B_h + B_v}{2}, A = \frac{A_h + A_v}{2}, Z = \frac{Z_h + Z_v}{2} \tag{10}$$

Finally the formulate an empirical model for the quality score:

$$Z_{score} = \alpha + \beta B^{\gamma_1} A^{\gamma_2} Z^{\gamma_3} \tag{11}$$

They used the non-linear regression routine available in the Matlab statistics toolbox to find the best value of parameters (α,β,γ_1,γ_2,γ_3) for Eq. 11. The values they calculated are used in the study presented here: $\alpha = -245.9$, $\beta = 261.9$, $\gamma_1 = -0.0024$, $\gamma_2 = 0.016$, $\gamma_3 = 0.0064$.

Blockiness is masked by the texture(spatial activity) in the region for which it is calculated. Activity measures are directly related to texture properties within blocks. The final Z-score is a nonlinear combination of these measures, that emulates the properties of the human visual system.

All the quality related features used are listed in Table 1. In the scope of the study presented here, they are good features to describe two regions of interest in test videos. Dynamic texture, which forms the salient part of the frame and the background which is non-salient.

It should be noted that both Wang *et al.* and Babu *et al.* measures were originally designed for 8×8 block size, which is the only size available in MPEG-2 [8], but not in MPEG-4/H.264/AVC [14]. However, since the size of the blocks in the latter case is constrained to 16×16, 8×8 or 4×4, the measures should be able detect blockiness and blurring along a subset of edges and within part of the blocks, and therefore can be used for any block-based codec.

Table 1 List of used quality features.

Salient reg. count	Z_{score} non-salient
Avg. reg. size	A salient
Mean change non-salient	B salient
Change Std.Dev. non-salient	Z salient
Mean Change salient	Z_{score} salient
Change Std.Dev. salient	BB non-salient
A non-salient	BB salient
B non-salient	BB border
Z non-salient	

4 Experiments and Results

4.1 Data Set

Videos from the DynTex data set [12] were used to evaluate the approach. The DynTex data set contains more than 650 varied dynamic texture videos, but the information about the type of textures shown in the sequences is not provided for all the videos in the set. Figure 1 shows example textures from this data set. The image size is 352×288 and the compressed videos provided were coded using DivX codec, i.e. an MPEG-4 Part 2 codec.

A subset of 202 sequences, spanning some 23 classes of very varied dynamic textures has been selected to evaluate the proposed approach. The subset is comprised of those DynTex sequences that were labelled as containing a single class of dynamic texture including those labelled NA for which the class information is not available. We treat the NA sequences as an additional class, increasing the diversity of the test set. The texture classes contained in our data set are: textile, vegetation, grass, NA, streaks, water, steam, fire, smoke, branch, cloud, leaf, car, flower, needle, fur, fish, tentacle, insects, CD, foam, light and paper. Sample frames from some of the sequences are shown in Figure 1.

4.2 Classification

Once the video signatures have been extracted, we used the Wakaito Environment for Knowledge Discovery (WEKA) tool [22] to train and test our classifiers. WEKA is an open source data mining and machine learning environment, which allows for different machine learning algorithms to be tested on a data set. The evaluation procedure conducted in two phases.

In the first phase we experimented with various algorithms, including the MLP suggested by Culibrk *et al.*, various decision trees and AdaBoost based on different algorithms. To determine the most suitable classifier, they were tested using the 10 fold stratified cross-validation methodology, i.e. 10% of data was withheld during training and used for testing. The data was selected randomly, but care was taken to preserve the distribution of the classes that exists in the original data set, since not all of the classes were represented with the same number of sequences.

In the second phase the best algorithm identified during the first stage was tested by withholding the VS of a single sequence for testing. The rest of the data was used for training. Once the classification for each part of the test VS was done, the mode of the VS was used to assign a class label to the sequence itself.

In our experiments, AdaBoost-ing the MLP classifier achieved the best result. In the first phase of the experiments, the classifier achieved 96.4% correct classification of each video signature part when tested on the training set and a 87.87% accuracy when cross-validated. The confusion matrix obtained for VS-part classification is shown in Figure 4 In our experiment we used AdaBoost M1 method, with 10 iterations. The MLPs contained 100 neurons, were trained using back-propagation with the learning rate of 0.3, moment 0.2 and 500 epochs.

In the second phase, since we used the mode of 10 classification values obtained per VS as the final label of the video, the cross-validated performance was nearly perfect (99.5% accurate) at the video-sequence level. The classifier failed to classify a single sequence accurately, out of 202 sequences in the data set. It should be noted that the sequence that the classifier failed to classify was one of 15 sequences in the database that are missing class information and is therefore quite possible that the training set contained no information that would enable the classifier to learn that particular case.

5 Conclusion

Although content-based video retrieval and indexing are usually stated as important potential applications of dynamic texture classification and recognition methodologies, there has so far been no attempt to address the problem of classifying dynamic textures based on compression artifacts.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	← classified as
195	0	0	3	0	13	1	0	1	0	3	0	0	0	0	0	0	6	0	1	1	6	0	a = textile
1	98	1	0	0	2	0	0	1	0	0	1	0	0	1	0	0	4	0	0	0	1	0	b = vegetation
0	0	53	0	0	6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	c = grass
4	1	0	128	0	6	0	0	0	1	1	0	1	0	0	0	0	7	0	0	0	1	0	d = ?
0	0	0	0	19	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	e = streaks
6	2	5	2	0	583	1	0	2	1	6	0	2	1	0	0	6	5	0	0	7	1	0	f = water
2	0	0	0	0	1	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	g = steam
6	0	0	0	0	1	1	10	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	h = fire
2	0	0	0	0	2	0	0	31	0	1	1	0	1	0	0	2	0	0	0	0	0	0	i = smoke
1	0	0	1	0	1	0	0	0	97	0	0	0	0	0	0	0	0	0	0	0	0	0	j = branch
0	0	0	0	0	5	1	0	0	0	54	0	0	0	0	0	0	0	0	0	0	0	0	k = cloud
0	0	0	0	0	1	0	0	0	0	0	8	0	0	0	0	0	1	0	0	0	0	0	l = leaf
1	0	0	1	0	2	0	0	2	0	0	0	44	0	0	0	0	0	0	0	0	0	0	m = car
2	1	0	0	0	7	0	1	0	1	0	0	0	33	0	0	0	4	0	0	0	0	1	n = flower
0	0	0	0	0	2	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	o = needle
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	8	0	1	0	0	0	0	0	p = fur
0	3	0	1	0	3	0	0	1	0	0	0	0	1	0	0	35	6	0	0	0	0	0	q = fish
4	3	3	0	1	7	0	0	0	1	0	0	0	0	0	1	1	149	0	0	0	0	0	r = tentacle
0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	7	0	0	0	0	s = insects
0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	27	0	0	0	t = CD
0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	u = foam
3	1	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	102	0	v = light
1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	w = paper

Fig. 4 Confusion matrix for the proposed classifier.

Since the performance of coding algorithms in terms of resulting video quality is profoundly dependent on the content of video, a novel approach to dynamic texture classification, which exploits this link is proposed in the paper. Specifically we showed that features commonly used for video quality assessment can be used, efficiently, to discern between different dynamic textures. The assumption made is that the videos are coded using the same codec and same bit-rates, which is not unreasonable in case of large multimedia databases.

An MLP-based AdaBoost classifier has been trained and evaluated using video quality features obtained through a state-of-the-art video quality assessment approach. A standard set of compressed dynamic texture videos has been used to test the approach. The approach achieves nearly perfect classification (99.5%) when cross-validated.

Several venues should be explored for further studies. The approach should be tested using other available databases, such as the UCLA database [15]. More importantly, the approach should be tested for different codecs. The blockiness and blurring measures designed may have to be adapted to handle the variable block-size of state-of-the art codecs explicitly.

Acknowledgements This research has in part been financed by the COST action IC0702 (SoftStat).

References

1. Babu V, Andrew P, Inge HO (2008) Evaluation and monitoring of video quality for UMA enabled video streaming systems. *Multimedia Tools Appl.* 37(2):211–231
2. Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32

3. Chan A, Vasconcelos N (2007) Classifying video with kernel dynamic textures. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1–6. IEEE Press, Piscataway
4. Chetverikov D, Péteri R (2005) A brief survey of dynamic texture description and recognition. *Computer Recognition Systems* 17–26
5. Culibrk D, Mirkovic M, Zlokolica V, Pokric M, Crnojevic V, Kukolj D (2010) Salient Motion Features for Video Quality Assessment. *IEEE Trans. on Image Processing* 948–958
6. Derpanis K, Wildes R (2010) Dynamic texture recognition based on distributions of spacetime oriented structure. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 191–198. IEEE Press, Piscataway
7. Ferzli R, Karam L (2007) A no-reference objective image sharpness metric based on just-noticeable blur and probability summation. *Proc. IEEE Conf. on Image Processing (ICIP)*, III:445–448. IEEE Press, Piscataway
8. Haskell B, Puri A, Netravali A (1997) *Digital video: an introduction to MPEG-2.* Kluwer, Amsterdam, Netherlands
9. Haykin S (1994) *Neural Networks: A Comprehensive Foundation.* Macmillan, New York
10. Kim K, Davis L (2004) A fine-structure image/video quality measure using local statistics. *Proc. IEEE Conf. on Image Processing* V:3535–3538
11. Kirenko I (2006) Reduction of coding artifacts using chrominance and luminance spatial analysis. *Proc. Int. Conf. on Consumer Electronics (ICCE)*, 209–210
12. Péteri R, Fazekas S, Huiskes M (2010) DynTex: A comprehensive database of Dynamic Textures. *Pattern Recognition Letters*
13. Ravichandran A, Chaudhry R, Vidal R (2009) View-invariant dynamic texture recognition using a bag of dynamical systems. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1651–1657. IEEE Press, Piscataway
14. Richardson I (2003) *H.264 and MPEG-4 video compression.* Wiley Online Library
15. Saisan P, Doretto G, Wu Y, Soatto S (2001) *Dynamic texture recognition*
16. Schapire R (2003) The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*, 149–172 Springer, New York
17. Schwenk H, Bengio Y (2000) Boosting neural networks. *Neural Computation* 12(8):1869–1887
18. Wang Z, Lu L, Bovik A (2004) Video quality assessment based on structural distortion measurement. *Signal processing: Image communication* 19(2):121–132
19. Wang Z, Sheikh HR, Bovik AC (2002) No-reference perceptual quality assessment of jpeg compressed images. *Proc. IEEE Int. Conf. on Image Processing*, 477–480
20. Warwick G, Thong N (2004) *Signal Processing for Telecommunications and Multimedia*, Chapter 6: Classification of Video Sequences in MPEG Domain. Springer, New York
21. Winkler S (2005) *Digital video quality: vision models and metrics.* J. Wiley & Sons, Chichester, United Kingdom
22. Witten IH, Frank E (2005) *Data Mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, San Francisco
23. Wolf S, Pinson M (1999) Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system. *Proc. Int. Symp. on Voice, Video, and Data Communications.* SPIE, Boston
24. YouTube: http://www.youtube.com/
25. Zechner M, Granitzer M (2009) Accelerating k-means on the graphics processor via CUDA. *Proc. Int. Conf. on Intensive Applications and Services (INTENSIVE 2009).* IEEE Press, Piscataway
26. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 915–928

The Hubness Phenomenon: Fact or Artifact?

Thomas Low[1], Christian Borgelt[2], Sebastian Stober[1], and Andreas Nürnberger[1]

Abstract The hubness phenomenon, as it was recently described, consists in the observation that for increasing dimensionality of a data set the distribution of the number of times a data point occurs among the k nearest neighbors of other data points becomes increasingly skewed to the right. As a consequence, so-called *hubs* emerge, that is, data points that appear in the lists of the k nearest neighbors of other data points much more often than others. In this paper we challenge the hypothesis that the hubness phenomenon is an effect of the dimensionality of the data set and provide evidence that it is rather a boundary effect or, more generally, an effect of a density gradient. As such, it may be seen as an artifact that results from the process in which the data is generated that is used to demonstrate this phenomenon. We report experiments showing that the hubness phenomenon need not occur in high-dimensional data and can be made to occur in low-dimensional data.

1 Introduction

That working with high-dimensional data is difficult has been known for quite some time now, although not all of the effects of a large number of dimensions are completely understood yet. In 1961, R.E. Bellman was among the first who recognized the various problems that arise in high-dimensional spaces, for which he coined the term *curse of dimensionality* [1]. One property of this "curse" is the fact that with an increasing number of dimensions the volume of a unit hyperball grows considerably less quickly than the volume of a unit hypercube. As a consequence, most distance metrics, like the Euclidean

[1] Data and Knowledge Engineering Group, Otto-von-Guericke-University of Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany, `firstname.lastname@ovgu.de`
[2] European Centre for Soft Computing, c/ Gonzalo Gutiérrez Quirós s/n, E-33600 Mieres (Asturias), Spain, `christian.borgelt@softcomputing.es`

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 267–278.
springerlink.com

distance, suffer from a loss of relative contrast. This effect is also known as *distance concentration* [2] and causes the relative difference in the distance of a given query point to its nearest and its farthest neighbor to vanish.

Another (alleged) property of the curse of dimensionality is the *emergence of hubs*, which was first described as a general problem in [10]. Here *hubs* are defined as data points that appear unusually often among the k nearest neighbors of other data points. Described in statistical terms, the distribution of the number of times a data point occurs in the nearest neighbor lists of other data points becomes skewed to the right. This phenomenon has been examined and demonstrated to be present in many real-world data sets in [11], where it was also analyzed how it affects a broad spectrum of machine learning tasks and dimensionality reduction techniques. The core claim of both papers, [10] and [11], is that the emergence of hubs is an intrinsic effect of the dimensionality of the data—a view we dare to challenge here.

Our core claim in this paper is that the hubness phenomenon is an effect of a density gradient, not an effect of the dimensionality of the data space. Note, however, that if the data points are sampled from a region that is bounded, there is necessarily a density gradient at the boundary of the region. Since the ratio of the size of the (hyper-)surface (i.e. the boundary) of a region to its (hyper-)volume increases exponentially with the dimensionality of the data space, the density gradient at the boundary is emphasized (in the sense that it influences more data points) and thus makes the hubness phenomenon more notable. This explains the observations of [11]. However, high dimensionality alone is not sufficient to produce the hubness phenomenon as we demonstrate by sampling from a boundary-less high-dimensional space, for which the hubness phenomenon is essentially nonexistent. We also show that the strength of the hubness phenomenon depends on the relative size of the boundary of the sampling region. In addition, we show that introducing sufficiently many boundaries (and thus many places with a density gradient) in a low-dimensional sampling region creates the hubness phenomenon.

The remainder of this paper is organized as follows: in Section 2 we define the hubness phenomenon and several measures by which we try to capture its strength, thus obtaining proper means to quantify and compare this phenomenon over different data sets. In Section 3 we describe our data generator, that is, the procedures we employed to generate high-dimensional data sets as well as the structure of these data sets. In Section 4 we describe the experiments (on artificial data) we conducted and report and interpret our results. Finally, in Section 5 we draw conclusions from our discussion.

2 Measuring Hubness

Before we can define measures for the strength of the hubness phenomenon, we have to introduce the notions on which it is based. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

be an m-dimensional data set with n data points $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ for $i \in \{1, \ldots, n\}$. Furthermore, let $N_k(\mathbf{x}) \subseteq \mathbf{X} - \{\mathbf{x}\}$ be the set of the k nearest neighbors ($k < n$) of $\mathbf{x}$ in $\mathbf{X}$. That is, $\forall \mathbf{y} \in N_k(\mathbf{x}) : \forall \mathbf{z} \in \mathbf{X} - N_k(\mathbf{x}) - \{\mathbf{x}\} : d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z})$ and $|N_k(\mathbf{x})| = k$ (assuming that ties are broken arbitrarily). We consider mainly the Euclidean distance $d(\mathbf{x}, \mathbf{y}) = (\sum_{j=1}^{n}(x_j - y_j)^2)^{\frac{1}{2}}$, but in principle other distance measures may also be studied (cf. [11]).

The quantity $o_k(\mathbf{x}) = \sum_{\mathbf{y} \in \mathbf{X}} \mathbb{1}_{N_k(\mathbf{y})}(\mathbf{x})$, where $\mathbb{1}_{N_k(\mathbf{y})}$ is the indicator function of $N_k(\mathbf{y})$ w.r.t. $\mathbf{X}$ (that is, $\mathbb{1}_{N_k(\mathbf{y})}(\mathbf{x}) = 1$ if $\mathbf{x} \in N_k(\mathbf{y})$ and 0 otherwise), counts the number of times the data point $\mathbf{x}$ occurs in the sets of nearest neighbors of other data points.[1] We call $o_k(\mathbf{x})$ the *k-occurrence* of the data point $\mathbf{x} \in \mathbf{X}$. The hubness phenomenon can now be described as the observation that the distribution of the values $o_k(\mathbf{x})$ for $\mathbf{x} \in \mathbf{X}$ is (considerably) skewed to the right or that some data points have unusually high k-occurrence values (i.e., considerably larger than the mean value, which is obviously k).

In order to obtain an objective evaluation of the strength of the hubness phenomenon, we rely on a few very simple measures. The most straightforward approach is obviously to compute the *skewness* (or simply *skew*) of the distribution of the $o_k(\mathbf{x}_i)$, $i \in \{1, \ldots, n\}$, which is defined as

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(o_k(\mathbf{x}_i) - \overline{o_k})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(o_k(\mathbf{x}_i) - \overline{o_k})^2\right)^{3/2}},$$

where $\overline{o_k} = k$ is the mean of the k-occurrence values (see [4] for a general definition of skewness). If $\gamma > 0$, the distribution is skewed to the right. However, usually only (very) few of the points in a data set are hubs, that is, exhibit high k-occurrence. Thus the skewness may not always be sensitive enough to measure the phenomenon properly. On the other hand, the skewness is computed from all data points and thus large values may result even if there are no sizable hubs, namely if there is an asymmetry close to the mean.

An alternative approach that immediately suggests itself is to use the largest k-occurrence in the data set. However, this measure has the disadvantage that it is strongly affected by the randomness of the sampling process. Thus it is only sufficiently expressive if averaged over a certain number of runs. To obtain a better measure we average the k-occurrence of the fraction q of data points with the highest k-occurrences, where q should be small. The averaging makes the measure more robust, yet allows us to properly capture the value of the highest k-occurrences.[2] In our experiments we tried $q = 0.1\%$, $q = 0.5\%$ and $q = 1\%$. Since this measure depends directly on the number k of nearest neighbors that are considered, we finally divide by k:

[1] Note that the sum need not exclude the data point $\mathbf{x}$, because $\mathbf{x} \notin N_k(\mathbf{x})$ by definition.

[2] We refrained from using the $(1 - q)$-quantile (which would be a an even more robust choice) because of the integer nature of the k-occurrences, which limits the number of possible values, especially for small k (that is, for few nearest neighbors). In addition, the $(1 - q)$-quantile does not capture the distribution of values at and beyond it.

$$h_1(q) = \frac{1}{k|O_k(q)|} \sum_{\mathbf{x} \in O_k(q)} o_k(\mathbf{x}),$$

where $O_k(q)$ contains the $\lfloor qn \rfloor$ data points with the highest k-occurrences. Note that this measure captures the maximum k-occurrence for $q = 1/n$.

As an alternative, we consider what percentage of the data points have a k-occurrence value at least β times the mean value k, that is, the percentage of data points $\mathbf{x} \in \mathbf{X}$ with $o_k(\mathbf{x}) \geq \beta k$. Formally we have

$$h_2(\beta) = \frac{|\{\mathbf{x} \in \mathbf{X} \mid o_k(\mathbf{x}) \geq \beta k\}|}{|\mathbf{X}|} \cdot 100\%.$$

In particular, we experimented with $\beta = 2$, $\beta = 3$ and $\beta = 4$. That is, if we consider, for example, the 10 nearest neighbors, we compute what percentage of the data points occurs in at least 20, 30, and 40 nearest neighbor lists. Note that h_2 highlights the number of hubs, while h_1 focuses on their size.

3 Data Generation

The design of a data generator starts with the choice of a (pseudo-)random number generator (RNG), usually for a uniform distribution. Here we rely on a simple and very fast RNG producing 32 bit unsigned integer numbers, which was suggested by G. Marsaglia [6]. This RNG computes the next (pseudo-) random number from the previous five numbers, has a period of about 2^{160}, and seems to pass all standard quality tests for RNGs. We prefer this RNG over the more fashionable Mersenne Twister [8] due to its simplicity and much higher speed. We use this RNG to generate uniformly distributed (pseudo-) random floating point numbers in the interval $[0, 1)$ by generating two 32 bit unsigned integers i_1 and i_2 and then computing $r = i_1 \cdot 2^{-32} + i_2 \cdot 2^{-64}$, thus filling all bits of the mantissa of a double precision floating point number. We ensure that $r \in [0, 1)$ by rejecting r and generating a new random number should the (highly unlikely) event occur that (due to rounding) $r = 1$.

In order to obtain normally distributed (pseudo-)random numbers (which we also need for sampling from (hyper-)spheres and (hyper-)balls, see below), we employ the so-called polar method [7, 5], which consists in generating two (pseudo-)random numbers x and y that are uniformly distributed in $[-1, 1)$ until $s = x^2 + y^2 < 1$, that is, until the point (x, y) lies inside a unit circle. Then the transformed numbers $x' = \xi x$ and $y' = \xi y$, where $\xi = \sqrt{-2 \ln(s)/s}$, are normally distributed with mean 0 and variance 1.

Our data generator can sample from a multivariate standard normal distribution as well as uniformly from a (hyper-)cube, a (hyper-)ball, and a (hyper-)sphere (i.e., the surface of a (hyper-)ball). Sampling uniformly from an m-dimensional (hyper-)cube $[-1, 1)^m$ (also called an m-cube) is trivial:

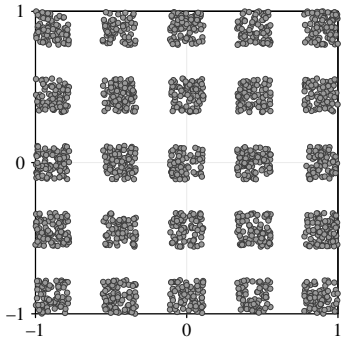


Fig. 1 A data set with 2000 points that were sampled uniformly from a grid of small squares described by the parameters $\eta = 5$ and $\alpha = 0.5$ (see the explanations in the main text).

simply generate m random coordinates $x_1, \ldots, x_m$ that are uniformly distributed in $[0, 1)$ and transform them with $x'_i = 2x_i - 1$ for $i = 1, \ldots, m$. Sampling from an m-dimensional standard normal distribution is equally simple: use the polar method (see above) to generate m normally distributed coordinates. The coordinates are combined in a vector $\mathbf{x} = (x_1, \ldots, x_m)$.

For sampling uniformly from an m-dimensional (hyper-)sphere (also called an m-sphere) we exploit the insight that a multivariate standard normal distribution is spherically symmetrical. Therefore, if $x_1, \ldots, x_m$ are sampled independently from a standard normal distribution, the vector $\mathbf{x}' = \mathbf{x}/||\mathbf{x}||$, where $\mathbf{x} = (x_1, \ldots, x_m)$ and $||\mathbf{x}|| = (\sum_{i=1}^{m} x_i^2)^{\frac{1}{2}}$, is uniformly distributed on an m-sphere [12]. In order to obtain points that are uniformly distributed over an m-dimensional (hyper-)ball (also called an m-ball), we start by sampling a vector $\mathbf{x}'$ uniformly from an m-sphere (see above) and in addition generate a (pseudo-)random number u that is uniformly distributed in $[0, 1)$. Since the radius r of a random vector that is uniformly distributed over an m-ball satisfies $P(r \leq z) = z^m$, we can write $r = u^{1/m}$. Therefore the vector $\mathbf{x}'' = r\mathbf{x}'$ is uniformly distributed over an m-ball [12].

In addition to these basic data generation modes, our implementation supports sampling uniformly from a regular grid of small (hyper-)cubes with a user-specified size. With this method we try to obtain a low-dimensional space with a large boundary in order to show that the hubness phenomenon can be produced in this way as well. The procedure is essentially the same as for sampling uniformly from a hypercube, only that the hypercube is cut into the requested number of small cubes and gaps are introduced by an appropriate transformation of the coordinates. To be precise: given a number η of (hyper-)cubes per dimension and a fraction α, which specifies how much (per dimension) of a grid cell is covered by the small (hyper-)cubes, we sample m coordinates $x_1, \ldots, x_m$ uniformly from $[0, 1)$ and transform them according to $x'_i = 2(\lfloor \eta x_i \rfloor + \alpha(\eta x_i - \lfloor \eta x_i \rfloor))/(\eta - 1 + \alpha) - 1$ for $i = 1, \ldots, m$. An example of such a 2-dimensional grid-structured data set with 2000 points, which was generated with $\eta = 5$ and $\alpha = 0.5$, is shown in Figure 1.

Finally, our implementation supports jolting points that were sampled uniformly from a (hyper-)cube into a (hyper-)ball. Intuitively, this can be seen

as "pushing in" the corners of the hypercube. Technically, this is achieved as follows: let $\mathbf{x} = (x_1, \ldots, x_m)$ be a data point in a (hyper-)cube. We determine $z = \max_{i=1,\ldots,m} |x_i|$ as well as $||\mathbf{x}|| = (\sum_{i=1}^{m} x_i^2)^{\frac{1}{2}}$. Then $\mathbf{x}' = z\mathbf{x}/||\mathbf{x}||$ lies inside a (hyper-)sphere with radius 1. Of course, the distribution of the resulting points is not uniform in the (hyper-)sphere, but denser near the axes from the center towards the corners of the original (hyper-)cube. However, it is very interesting to see how this transformation affects the hubness phenomenon, as the result is not quite what one might expect.

4 Experimental Results

With the four experiments we describe in the following we try to clarify the inherent properties of the hubness phenomenon. In the first experiment we show that hubness need not occur in high-dimensional spaces by sampling from a finite, but boundary-less space. We demonstrate that it is rather directly related to a density gradient. In the second experiment we show that hubs also occur in low-dimensional spaces and reveal the true cause of the hubness phenomenon. The third experiment demonstrates the dependence of the hubness phenomenon on artificially introduced density gradients, especially the size of the surface of the sampling region. Finally, the fourth experiment examines the effect of jolting a (hyper-)cube into a (hyper-)ball.

Experiment 1: As already reported in [11], the distribution of the k-occurrences becomes skewed to the right if a data set is sampled uniformly from a high-dimensional hypercube, and even more so if the data set is sampled from a high-dimensional normal distribution. Our experiments confirm this observation, as can be seen from the curves labeled "cube" and "normal" in Figure 2.[3] For $m \geq 15$, and certainly for $m \geq 20$, all three hubness measures (skewness γ, $h_1(1\%)$, i.e. the average k-occurrence, divided by k, of the data points with the 1% highest k-occurrences, and $h_2(3)$, i.e. the percentage of data points occurring in a least $3k = 30$ nearest neighbor lists) clearly indicate a strongly skewed distribution and the existence of sizable hubs.

It is remarkable that data sampled from a normal distribution exhibit a much stronger hubness phenomenon than data sampled uniformly from a hypercube. This finding already provides a hint that the hubness phenomenon is caused by a density gradient, because a normal distribution possesses a strong density gradient everywhere, while data sampled from a hypercube possess such a gradient only at its hyperfaces. As these faces cover a considerable hyperarea for high-dimensional data, many data points are influenced by the gradient they cause, but still fewer than in a normal distribution.

[3] All such diagrams in this paper have been obtained by averaging over 200 runs in order to reduce the effects of randomness and to achieve representative results.

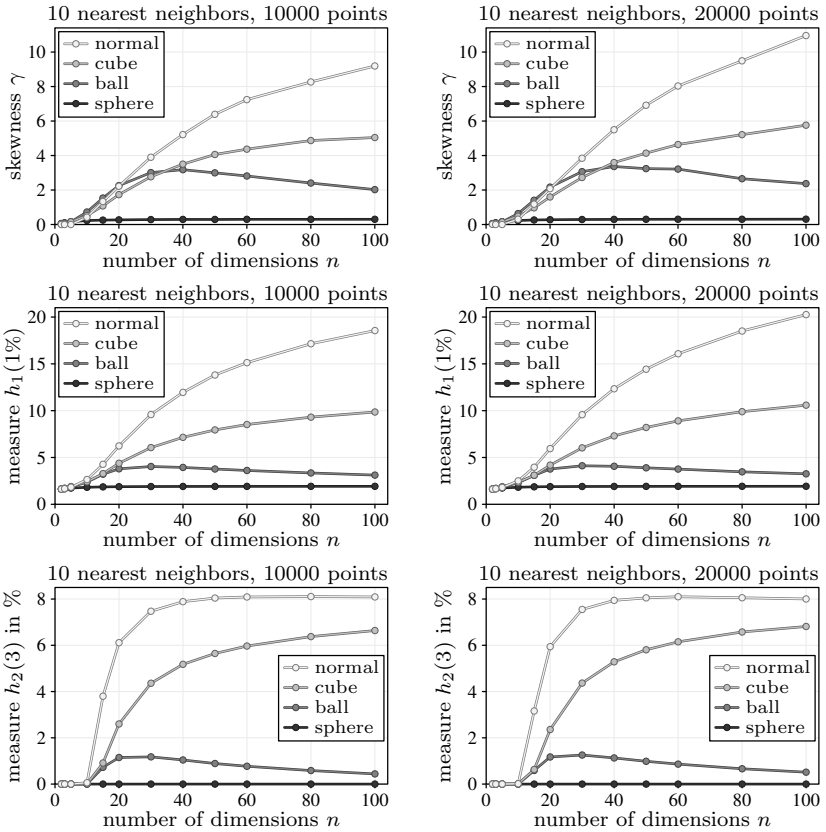


Fig. 2 The hubness phenomenon for different sampling distributions (multivariate normal and uniformly from a (hyper-)cube, (hyper-)ball and (hyper-)sphere) and two data set sizes, assessed by different hubness measures (as defined in Section 2).

A further hint is provided by the fact that data sampled uniformly from a hyperball (see curves labeled "ball" in Figure 2) exhibit a much lesser hubness phenomenon, which even is reduced again beyond $m \approx 30$–40. Since the sample is still drawn uniformly, the different strength of the hubness phenomenon must be due to the different shape of the sampling region. We believe that the absence of "corners" (at which the density gradient is particularly high) and the much smaller hypersurface relative to the enclosed hypervolume are the reason for this effect. Since the hypersurface of a hyperball is much smaller compared to that of a hypercube, fewer data points are affected.

However, the strongest argument that high dimensionality alone does not cause a hubness effect is provided by the following consideration: an $(m+1)$-dimensional hypersphere is essentially an m-dimensional space, but finite and boundary-less. As a consequence, there is no density gradient anywhere, at least if we confine ourselves to the topology of the hypersphere (which is

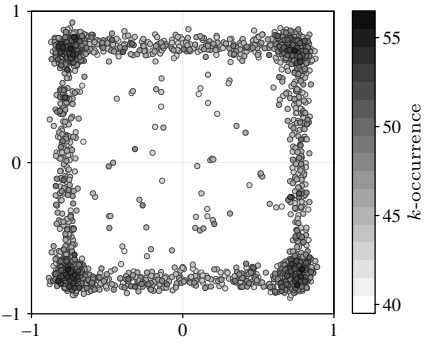


Fig. 3 The 12 largest hubs collected from 250 sets of 1000 points each, sampled uniformly from a square $[-1, 1)^2$. For each point the 30 nearest neighbors were considered. The darkness of a circle encodes a hub's k-occurrence.

equivalent to using an elliptic geometry). Although this does not pose any problems (the shortest path between two points follows the meridian through them and thus, for a unit hypersphere, its length is equal to the angle between the points, measured in radians), we rely on the Euclidean distance in the $(m + 1)$-dimensional space, which yields essentially the same result. As can be seen from the curves labeled "sphere" in Figure 2, none of the measures detects a hubness phenomenon, regardless of the number of dimensions.

Note that this finding explains why a (hyper-)ball exhibits a much less pronounced hubness effect: for increasing dimensionality the mean Euclidean norm of points sampled uniformly from a unit (hyper-)ball converges to 1 (see Section 3: $P(r \leq z) = z^m$). Thus we may say that in high-dimensional spaces almost all points in a (hyper-)ball lie close to its surface and thus almost on a (hyper-)sphere. With this regard it is no longer surprising that the hubness phenomenon reduces again for very high-dimensional (hyper-)balls.

Experiment 2: In our second experiment we show that the hubness phenomenon also occurs in low dimensional spaces, although not as pronounced, and reveal its true cause. In the diagram in Figure 3 the twelve largest hubs have been collected from 250 data sets that were sampled uniformly from a square. The darkness of the hubs encodes their k-occurrence, which shows that there are not only *more hubs* close to the corners, but that these hubs also tend to be *larger* (that is, they tend to have higher k-occurrences).

This effect is clearly due to the boundaries of the square and can be seen as a kind of mirroring effect. Points close to the sides and corners do not have as many options to choose their nearest neighbors compared to points in the interior of the square. Therefore points that lie near points that are close to the sides and corners are more likely to be chosen as nearest neighbors and thus become hubs. Hubs are more frequent close to the corners, because here up to 3/4 of the space (for a point exactly at the corner of square) are void of points, while in the middle of a side only up to 1/2 of the space is void of points. If one extrapolates this finding to more dimensions, it becomes clear why hypercubes exhibit such a strong hubness phenomenon.

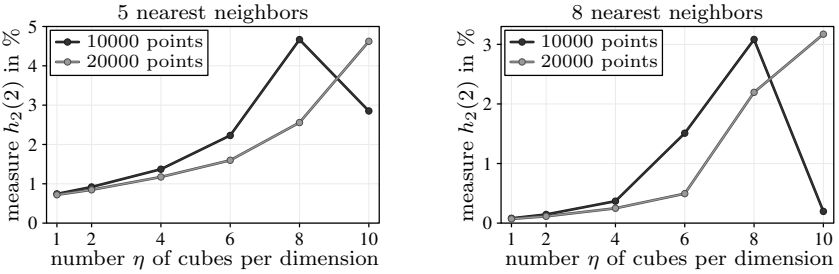


Fig. 4 Hubness in 3-dimensional data sets sampled uniformly from a grid of cubes: dependence on the number of cubes/grid cells per dimension (gap size $\alpha = 50\%$).

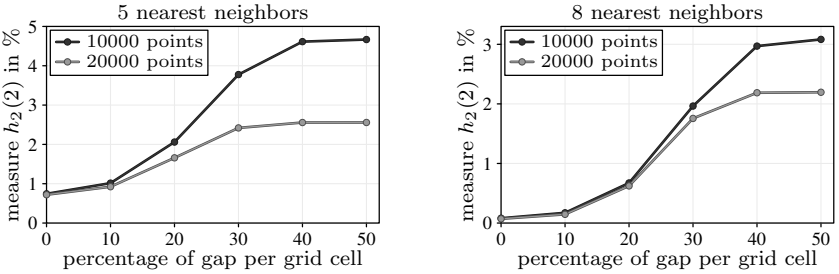


Fig. 5 Hubness in 3-dimensional data sets sampled uniformly from a grid of cubes ($\eta = 8$ cubes per dimension): dependence on the size of the gaps (measured per dimension).

Experiment 3: Our third experiment expands on our view that the size (and shape) of the (hyper-)surface (and the density gradient it causes) produces the hubness phenomenon. If this view is correct, it should be possible to create a hubness phenomenon in a low-dimensional space by sampling from a region with a large boundary. Our core idea is to sample data from a grid of squares or cubes. If the gaps between these squares or cubes are big enough, so that a nearest neighbor is almost surely found in the same cube, there should also be a certain, though weaker hubness phenomenon.

This effect is demonstrated in Figure 4, which shows how a grid of cubes (gap size $\alpha = 50\%$) leads to a hubness effect with an increasing number of cubes. The effect is weak, though, but can be detected with the skewness γ or with the measure $h_2(2)$ used in the diagrams. Note that the effect is generally bounded due to the topology of a 3-dimensional space, as can be seen from the relation of the hubness phenomenon to the kissing number problem [3, 9]. A kissing number is the number of non-overlapping unit spheres that touch another given unit sphere. As was already noticed in [11], this problem is relevant for the hubness phenomenon, because a data point cannot be the nearest neighbor of more data points than the kissing number of the space it resides in. Since the kissing number for three dimensions is 12, sizable hubs are extremely unlikely as they require highly symmetric point arrangements.

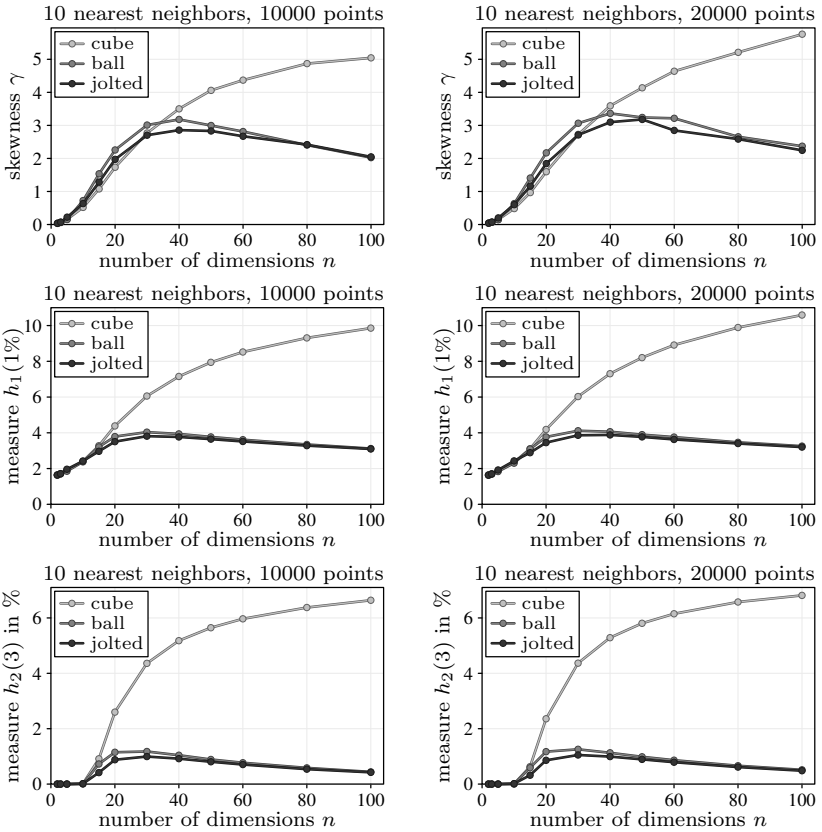


Fig. 6 The hubness phenomenon for data that was sampled uniformly from a (hyper-)cube and then jolted into a (hyper-)ball, as compared to sampling uniformly from a (hyper-)cube or (hyper-)ball, assessed with the same measures as in Figure 2.

Note that for 10000 points the hubness phenomenon disappears again for 10 cubes per dimension. The reason for this effect is simply that under these circumstances the average number of points per cube is 10 (as there are $10 \times 10 \times 10 = 1000$ cubes). If 8 nearest neighbors are considered, almost all points in a cube are the nearest neighbors of all other points in the same cube. As a consequence, the location of the points relative to the boundary, which is responsible for the hubness effect (see Experiment 2), becomes irrelevant.

Note also that the gaps between the cubes have to be large enough, as can be seen in Figure 5: for small gaps there is basically no hubness effect, because nearest neighbors may still be found in neighboring cubes, thus reducing or even eliminating the effect of the cube surfaces/boundaries.

Experiment 4: Our last experiment reveals a somewhat unexpected behavior that we discovered during our analysis. As we have seen in Experiment 1, the skewness of the distribution of k-occurrences for data sampled uniformly

from a (hyper-)cube is much stronger than for data sampled uniformly from a (hyper-)ball. This led to the idea to "jolt" a (hyper-)cube into a (hyper-)ball in order to check whether this operation reduces the hubness phenomenon. Given our view of the causes of the hubness phenomenon, we certainly expected it to be reduced (because this operation significantly reduces the surface of the sampling region), but that it was even reduced slightly below the level of data that was sampled uniformly from a (hyper-)ball was somewhat surprising (see Figure 6, curve labeled "jolted"). We rather expected it to lie between a (hyper-)cube and a (hyper-)ball based on the argument that the jolting introduces density gradients inside the (hyper-)ball.

However, on second thought, the effect becomes understandable. The jolting operation, even though it causes density gradients inside the (hyper-)ball, also reduces the effect of the (remaining) (hyper-)surface, because it pushes more data points into the interior of the (hyper-)ball, thus leaving less at the surface that cause the hubness effect (cf. Experiment 2).

5 Conclusions

In this paper we demonstrated that the hubness phenomenon is not an effect of the (high) dimensionality of a data set, but an effect of a density gradient. However, a density gradient may be intrinsic to the data set (if the data is not uniformly distributed) or it may be a boundary effect. Since a boundary necessarily introduces a density gradient and the ratio of the size of the boundary to the size of the enclosed space grows exponentially with the dimensionality of the data space, high-dimensional bounded data is prone to exhibit the hubness phenomenon. However, it is important to note that this phenomenon can also be produced, though much weaker, in a low-dimensional space by artificially increasing the size of the boundary. Another factor is the shape of the boundary: "corners" intensify the effect, as can be seen from the pronounced hubness phenomenon exhibited by (hyper-)cubes.

Software
Our implementation of the data generator and evaluation routines, which we used for the experiments in this paper as well as the corresponding Python scripts automating the experiments will soon be available for download at:
`http://www.borgelt.net/hubness.html`

Acknowledgements The work presented in this paper was supported by Short-Term Scientific Mission (STSM) grant 7684 (Thomas Low) of COST Action IC0702.

References

1. Bellman RE (1961) *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ, USA 1961
2. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is "Nearest Neighbor" Meaningful? *Proc. 7th Int. Conf. on Database Theory (ICDT 1999, Jerusalem, Israel)*, LNCS 1540:217–235. Springer-Verlag, Berlin, Germany
3. Conway JH, Sloane NJA (1999) *Sphere Packings, Lattices and Groups (3rd edition)*. Springer-Verlag, New York, NY, USA
4. Groeneveld RA, Meeden G (1984) Measuring Skewness and Kurtosis. *J. of the Royal Statistical Society, Series D (The Statistician)* 33(4):391399. Blackwell Publishing, Oxford, United Kingdom
5. Knuth DE (1998) *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*. Addison-Wesley, Reading, MA, USA
6. Marsaglia G (2003) Re: good C random number generator. Post on newsgroup `comp.lang.c`, date: 2003-05-13 08:55:05 PST. `http://groups.google.com/group/comp.lang.c/browse_thread/thread/a9915080a4424068/`
7. Marsaglia G, Bray TA (1964) A Convenient Method for Generating Normal Variables. *SIAM Review* 6:260-264. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA
8. Matsumoto M, Nishimura T (1998) Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudorandom Number Generator. *ACM Trans. on Modeling and Computer Simulation* 8:3-30. ACM Press, New York, NY, USA
9. Nebe G, Sloane NJA (2012) *Table of the Highest Kissing Numbers Presently Known*. `http://www.math.rwth-aachen.de/~Gabriele.Nebe/LATTICES/kiss.html` (retrieved 2012.01.16)
10. Radovanović M, Nanopoulos A, Ivanović M (2009) Nearest Neighbors in High-Dimensional Data: The Emergence and Influence of Hubs. *Proc. 26th Int. Conf. on Machine Learning (ICML 2009, Montreal, Canada)*, 865–872. ACM Press, New York, NY, USA
11. Radovanović M, Nanopoulos A, Ivanović M (2010) Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *J. Machine Learning Research* 11(Sep):2487–2531. MIT Press, Cambridge, MA, USA
12. Rubinstein RY, Kroese DP (2007) *Simulation and the Monte Carlo Method (2nd ed.)*. J. Wiley & Sons, Chichester, United Kingdom

Proximity-Based Reference Resolution to Improve Text Retrieval

Shima Gerani[1], Mostafa Keikha[1], and Fabio Crestani[1]

Abstract Queries that contain named entities are very common especially in the blog retrieval. Current approaches for document retrieval are based on the frequency of query terms in documents. These methods may underestimate the query frequency due to the fact that named entities are usually referenced using anaphoric expressions. In this paper we focus on pronouns as anaphoric expressions and propose a method for finding query-entity types including *female, male* and *non-person* which helps to identify the proper set of pronouns that can refer to each query. We also propose a proximity-based method for estimating the frequency of the anaphoric expressions which are referring to a query-entity in a document. Experimental results on a standard blog collection show that the proposed method is effective and provides significant improvement over the term-frequency-based baseline.

1 Introduction

Queries that contain named entities are very popular in search applications, especially in the context of Blog retrieval. One of the drawbacks of current approaches for entity retrieval is the underestimation of query-entity (i. e. the entity referred by the query). Frequency in the documents. This is due to the fact that named entities are usually referenced using anaphoric expressions such as pronouns. For instance, if a document contains a person entity, pronouns such as *he, him, his, etc.*, may be used instead of repeating the person's name. Considering anaphoric expressions that refer to the query-entity can help to better estimate the query occurrence in the documents and improve retrieval performance. The key problems are: 1) identifying a feasible set of anaphoric expressions for the query-entity since different pronouns are used

[1] Faculty of Informatics, University of Lugano, Switzerland,
{shima.gerani,mostafa.keikha,fabio.crestani}@usi.ch

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 279–290.
springerlink.com

to refer to person (female, male) or non-person entities, and 2) estimating the frequency of anaphoric expressions that are referring to the query-entity. A recent study [4] has highlighted this problem and proposed estimating Coreferentially Enhanced Entity Frequency (CEEF), to include the frequency of anaphoric expressions in the frequency count of query entities in documents. CEEF learns a Support Vector Machine(SVM) classifier with two features based on the standard frequency of anaphoric expressions in the top M relevant documents, in order to classify query entities to person/non-person types. The CEEF model assumes the presence of K non query-entities in every document and estimate the probability that all anaphoric expressions in a document are referring to the query-entity rather than other K entities. A 2-Poisson model is used to estimate this probability. The standard frequency of anaphoric expressions is then weighted by this probability and is added to the query-entity frequency. CEEF is based on the query-entity frequency rather than individual query term frequencies and involves manual identification of parts of the query which reflects the query-entity. CEEF also assumes the presence of a single entity in every query. For instance in queries such as *March Warner for president*, Mark Warner is the entity and the rest is removed manually. Another problem with entity phrase-based retrieval is that, they underestimate the query entity frequency in cases where the query entity is referred using a part of the phrase. For instance *Cindy Sheehan* may occur once at the beginning and be referred by just *Cindy* through the rest of the document.

In order to avoid problems such as preprocessing of a query to identify a query-entity phrase and also underestimation of the query entity, in this paper we do not limit ourselves to the entity phrase and follow the standard approach of term-based retrieval. In [4] after identifying the query entity phrase and counting it through the document, the entity frequency is enhanced by adding the weighted frequency of its feasible anaphoric expressions in the document as we explain in the following: First, all occurrences of the feasible anaphoric expressions in the document are counted. Then this anaphoric frequency is weighted by the probability that they are referring to the query entity and is added to the query-entity frequency. The probability that the anaphoric expressions are referring to the query-entity is the same for all feasible pronouns and is calculated based on the eliteness of query in the document using a 2-poisson model.

In this paper, we propose a method for finding query-entity types including *female, male* and *non-person* which helps us in identifying the proper set of pronouns for each query. We also propose a proximity-based method for estimating the frequency of the anaphoric expressions which coreference the query terms. Unlike CEEF, our proposed method does not consider all feasible pronoun sets equally and count them differently based on their proximity to the query terms. The proximity information has been used previously to find opinion terms relevant to the topic in the context of blog retrieval [2, 7, 10]. Here we use the proximity information to estimate the probability

that the pronoun is referring to the same entity as the query does. Experimental results on the BLOG06 collection of the TREC Blog Track show that the proposed method is effective and provides significant improvement over the raw term frequency baseline. Thus, our contributions are:

- Presenting a novel method for identifying the query entity type.
- Presenting a proximity-based method for estimating the probability that a specific anaphoric expression refers to the query-entity.
- Evaluating the method on a large collection and investigating the impact of the different components of the proposed model.

The remainder of this paper is organized as follows: In section 2 we explain and motivate the problem. Section 3 introduces our proposed method for estimating the real frequency of query in a document. Section 4 describes the retrieval method and section 5 reports on the experiments we conducted in order to evaluate the usefulness of the proposed model. Finally in Section 6, we conclude the paper and describe the future work.

2 Beyond Observed Term Frequency

The most important component of an information retrieval system for estimating the relevance of a document to a query is the query term frequency in the document. However, the query entity is usually referred to using anaphoric expressions instead of the exact query-entity name. Therefore, the exact counting of query terms underestimates the relevance of documents to the query entities. A better estimation of query term frequency in a document can be obtained by including the frequency of anaphoric expressions that coreference with the query term. Assume w to be a term in query Q. The real frequency of w can be calculated as follows:

$$tf(w;d) = tf_{observed}(w;d) + tf_{anaphoric}(w;d) \tag{1}$$

where $tf_{observed}(w;d)$ is the standard frequency of query term w in document d and $tf_{anaphoric}(w;d)$ is the amount of w occurrences evidenced by anaphoric expressions in d. If we assume that all anaphoric expressions are coreferential with the query terms, standard term frequency can be applied to calculate the frequency of anaphoric expressions as well. However, this assumption is very simplistic and it is quite possible that one or more entities of the same type as the query-entity exist in a document, and that some or all of the anaphoric expressions in the document refer to non-query entities. One may think of using coreference resolution systems [8] to identify the pronouns that are referring to the query entity. However, the accuracy of even state-of-the-art coreference resolution systems is less than 70%. The other issue is the efficiency of such systems. They are usually based on Natural Language

This is what I know about the events of January 31, 2006 involving **Cindy Sheehan**. **Cindy** not only didn't protest at the SotU, **she** didn't even plan - or want - to attend it. **Cindy** was set to... (omitted).
I was with **Cindy**, **her** sister Didi, and **her** friends Ann Wright and Bill Mitchell immediately before **she** went to the Capitol Building to watch the State of the Union address. I saw **Cindy** pass the ticket for the SotU to **her** friend Gael Murphy to give to Iraq War vet John Bruhns. We protested together outside the White House, where **Cindy** took several cell phone calls telling **her** the media were playing up Rep. Lynn Woolsey's invitation. Then **she** drank hot chocolate and ate baby carrots while we discussed whether or not she should change **her** mind and attend the speech....(omitted).

Fig. 1 An excerpt from blog post BLOG06-20060216-032-0006921029

Processing modules which take a lot of time and are not very efficient [4]. An example that has been discussed in previous studies is BART [9] which takes about 24 hours to perform coreference resolution of only 1,000 documents in the TREC Blog Track on an Intel 2.83GHz CPU. Considering the huge amount of documents in our applications (e. g. 3.2 million documents in Blog06 collection) exact coreference resolution is not practical.

In this paper we propose a method to approximate the frequency of anaphoric expressions that coreference an entity with a query term. Before explaining our model for estimating $tf_{anaphoric}(w; d)$, we need to introduce some notations and settings. We denote a document as a vector $d = (t_1, ..., t_i, ..., t_j, ..., t_{|d|})$ where the subscripts i and j indicate positions in the document and t_i indicates the term occurring at the position i. Accordingly, we denote a query with vector $Q = (w_1, ..., w_{|Q|})$. In this study, we limit the set of anaphoric expressions to pronouns. Assuming three types of query entity, *female person, male person* and *non-person*, we consider the following set of pronouns as possible references to each entity type:

- Male person: $A_{mp} = \{he, his, him, himself\}$
- Female person: $A_{fp} = \{she, her, herself\}$
- Non-person: $A_{np} = \{it, its\}$

An excerpt from a blog post in the BLOG06 collection is shown in Figure 1. As we can see, there are a lot of references to *Cindy Sheehan* using pronouns such as *her* or *she* which are ignored in case of using standard term frequency. This causes the underestimation of the query term frequency in the document. In the next section, we explain a proximity-based model for estimating the frequency of pronouns that coreference a set of query terms. We then use this model for estimating the anaphoric frequency of a single query term in section 3.2 and for classifying the query-entity type in section 3.1.

3 Proximity-based Coreference Resolution

The frequency of pronouns that are co-referenced with a set of query terms ($X \subseteq Q$) in document d, can be estimated based on the following three assumptions:

Assumption 1: Pronoun occurrences in a document d are counted, if all query terms occur at least once in d.

Assumption 2:A pronoun at position i in d can not refer to X that has not occurred yet in d ($\forall j \in pos(X), j > i$).

Assumption 3: A pronoun is more probable to be co-referenced with X if it occurs in close proximity after an occurrence of X in d.

If we consider A_k as the feasible set of pronouns for query Q, the anaphoric frequency of X can be estimated as follows:

$$atf_X(A_k; d) = \sum_{i=firstPos(X)+1}^{|d|} p(r_X|i)c(A_k; i, d), \tag{2}$$

where, "*firstPos(X)*" indicates the first position of any element of X in d. The value of $c(A_k; i, d)$ is 1 if the term at position i is a pronoun from set A_k and is 0 otherwise. $p(r_X|i)$ is the probability that an anaphoric expression at position i, coreferences X. We estimate this probability using a Gaussian function [1] as follows:

$$p(r_X|i) = \exp\left[\frac{-(j-i)^2}{2\sigma^2}\right] \qquad \text{where} \qquad j = \underset{j \in pos(X) < i}{\operatorname{argmin}} (i - j). \tag{3}$$

Gaussian function has a sigma parameter,σ, that here identifies the distance in which an anaphoric expression and a query term can be coreferential. This probability slowly decreases with the distance of the pronoun from the occurrence of w and is maximum (close to 1) at positions immediately after an occurrence of w. Note that according to our first assumption, the sum in Equation (2) starts from the first position of w in d, therefore, if X does not occur in d, none of the pronouns can be considered as coreferential with X and so $atf_X(A_k; d)$ remains zero.

Note that in the above formulas X can be a single or multiple terms from the query, which as a whole refer to an entity. Since identifying the set of entity-referring terms of a query, in an automatic way, may be challenging, we propose considering every term of the query separately, as X, and estimate its anaphoric counts which can then be added to the standard query term count and be used in the retrieval models. In the rest of this section we first explain how we use $p(r_X|i)$ in estimating the anaphoric-enhanced query term frequency and then we explain the usage of $p(r_X|i)$ for feasible pronoun set identification.

3.1 Query Type Identification

As a first step in our model we need to identify the query-entity type(s). We propose a feedback-based method, to rank pronoun sets based on the probability of being feasible for the query-entity. We calculate the average probability of each pronoun set in the top M relevant documents as follows:

$$f_k(Q) = \frac{1}{|F(Q)|} \sum_{d \in F(Q)} \frac{atf_Q(A_k; d)}{len(d)},$$

where $F(Q)$ is the set of top M relevant documents to the query Q and k indicates the query type which can be one of: mp, fp or np. The frequency of pronoun set A_k, which are coreferenced with Q, can be estimated using Equation 2. The probability that the anaphoric set A_k is feasible for query Q can then be calculated as follows:

$$P(A_k|Q) = \frac{f_k(Q)}{\sum_{k \in \{mp,fp,np\}} f_k(Q)} \tag{4}$$

3.2 Anaphoric Frequency of Query Terms

The anaphoric frequency of a query term can be estimated as follows:

$$tf_{anaphoric}(w; d) = \sum_{A_k \in A} P(A_k|Q) atf_w(A_k; d) \tag{5}$$

where A is the set of pronoun categories that are feasible for the query, that is $A = \{A_{mp}, A_{fp}, A_{np}\}$. $P(A_k|Q)$ is the probability that A_k is a feasible pronoun set for the query entities. Some queries may contain more than one entity type. For instance, query Q_{1009} from BLOG06 collection, *"Frank Gehry architecture"*, contains *Frank Gehry* as a person entity, but the whole query can refer to the *architecture* by Frank Gehry. Ideally we want to consider this ambiguity in the query-entity type and reflect it in the probability $P(A_k|Q)$ as follows: $P(A_k|Q_{1009})$ is 0.57, 0.41 and 0.02 for pronoun sets A_{np}, A_{mp} and A_{fp} respectively. These probabilities indicate that it is more probable that the query refers to a *male* entity and It is also highly probable that the query-entity is a non-person entity, but it is unlikely that the query is referring to a *female* entity. After ranking pronoun sets based on $P(A_k|Q)$, we may choose to use the top one, two or all pronoun sets. In the next section, we describe a feedback proximity-based method to calculate this probability for every pronoun set in relation to the query.

$atf_w(A_k; d)$ is an estimation of the number of times that a pronoun from A_k is coreferential with query term w in document d and can be estimated using Equation 2 where $X = \{w\}$.

Once we obtain the new query term frequency, using Equation (1), estimated based on observed and anaphoric frequencies, we can apply any retrieval model to rank documents according to their relevance to the query.

4 Experiments

In this section we first explain the experimental setup used for evaluating our methods. We then evaluate our proposed method for identifying the query type such as male, female or object in Section 4.2. We report and discuss the result of evaluating our method on test queries in Section 4.3.

4.1 Experimental Setup

Our experiments are based on the BLOG06 collection [3] which contains more than 3 million blog posts. We used the set of 150 topics in TREC 2006 through 2008 and their corresponding relevance assessments.

Each permalink was preprocessed by removing boiler templates (i. e. non-relevant parts such as menu, banner, site description, etc) using DiffPost algorithm [5] and then indexed as a retrieval unit. The preprocessing of the collection was minimal and involved only stopword removal with pronouns removed from the stop list. To set the parameters of the model (i.e., σ, okapi parameters, and the number of feedback documents), we used TREC06 ($Q_{851} - Q_{900}$) topics as our training set and TREC07 ($Q_{901} - Q_{950}$) and TREC08 ($Q_{1001} - Q_{1050}$) topics for our test set.

We used the topical relevance judgements provided by TREC for evaluation. We report the MAP as well as R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10). Throughout our experiments we used the paired t-test with significance level of 0.001 to test the significance of our method. The symbols $-$ and $+$ show statistical significant decreases and increases compared to the standard term frequency (STF) baseline.

In this paper we use Okapi BM25 to score documents [6]. The BM25 model is explained as:

$$BM25 = \sum_{w \in Q} idf(w) \frac{(k_1 + 1))tf(w; d)}{k_1((1 - b) + b(l_d / L_{ave})) + tf(w; d)}$$

Here, k_1 and b are normalization parameters. L_{ave} indicates the average document length in the collection and $idf(w)$ is the inverse document frequency

QNo	Query	Query type weighting	Evaluation
864	muhammad cartoon	$P(A_{mp}\|Q)$:0.57, $P(A_{fp}\|Q)$: 0.07, $P(A_{np}\|Q)$: 0.36	Correct
869	colbert report	$P(A_{mp}\|Q)$:0.55, $P(A_{fp}\|Q)$: 0.03, $P(A_{np}\|Q)$: 0.42	Correct
935	Mozart	$P(A_{mp}\|Q)$:0.74, $P(A_{fp}\|Q)$: 0.01, $P(A_{np}\|Q)$: 0.25	Correct
1006	Mark Warner for President	$P(A_{mp}\|Q)$:0.68, $P(A_{fp}\|Q)$: 0.15, $P(A_{np}\|Q)$: 0.17	Correct
1019	China one child law	$P(A_{mp}\|Q)$: 0.07, $P(A_{fp}\|Q)$: 0.20,$P(A_{np}\|Q)$:0.73	Correct
1049	Aperto Networks	$P(A_{mp}\|Q)$: 0.00, $P(A_{fp}\|Q)$:0.00,$P(A_{np}\|Q)$:1.0	Correct
905	king funeral	$P(A_{mp}\|Q)$:0.61, $P(A_{fp}\|Q)$: 0.29, $P(A_{np}\|Q)$: 0.1	Incorrect
1034	Ruth Rendell	$P(A_{mp}\|Q)$:0.59, $P(A_{fp}\|Q)$: 0.21, $P(A_{np}\|Q)$: 0.20	Incorrect

Table 1 Example of pronoun set weighting for queries

of w:

$$idf(w) = \log(\frac{N - n + 0.5}{n + 0.5})$$

4.2 Evaluation of Query Type Identification

To evaluate the accuracy of our query type identification method, we manually classify the queries to the three categories of *male, female* and *non-person*. We then compare the manually assigned type to the first category of pronouns assigned to every query of test set. The comparison shows that the proposed method has classification accuracy of 0.98. Out of 100 test queries, just two queries, Q_{905} and Q_{1034}, were assigned an incorrect primary pronoun set. Table 1 shows examples of correct and incorrect entity-query type weighting. The weights and rankings of pronoun sets, perfectly reflects the ambiguity of some queries such as Q_{1019}, *China One Child Law*, where we can consider the whole query phrase as a non-person entity or we can consider *child* as a person entity. As we can see from Table 1, the probability assigned to the non-person entity is more than the probability of female or male type, which makes sense. On the other hand, in cases such as Q_{1049}, *Aperto Networks*, the query is non-person with no ambiguity and it is reflected in the weighting clearly.

4.3 Retrieval Performance on the Test Queries

Table 2 shows the evaluation result of our proposed method on the test set of queries. PW indicates the proposed method for estimating anaphoric frequency of query terms using Equation 5. We should remind that the anaphoric frequency is then added to the standard term frequency of the query term using Equation (1).

method	MAP	R-prec	bPref	p@10
STF	0.4152	0.4606	0.4933	0.6710
T_1	0.3109^-	0.3665^-	0.4526	0.4170^-
T_2	0.2908^-	0.3470^-	0.4339^-	0.3720^-
T_3	0.2812^-	0.3390^-	0.4267^-	0.3350^-
W_1	0.3269^-	0.3804^-	0.4632	0.4590^-
W_2	0.3211^-	0.3728^-	0.4572	0.4450^-
W_3	0.3194^-	0.3717^-	0.4555	0.4450^-
P_1	0.4319^+	0.4683	0.5230^+	0.7070
P_2	0.4282^+	0.4703	0.5243^+	0.7330^+
P_3	0.4283^+	0.4709	0.5207^+	0.7270^+
PW_1	0.4365^+	0.4685^+	0.5206^+	0.7140^+
PW_2	0.4397^+	0.4731^+	0.5228^+	0.7180^+
PW_3	0.4402^+	0.4727^+	0.5233^+	0.7240^+

Table 2 Results on TREC07-08 query sets.

In order to see the effect of different components, we consider the following variations of the model:

Variation T

Variation T estimates $tf_{anaphoric}(w;d)$ using the standard frequency of pronouns:

$$tf_{anaphoric}(w;d) = \sum_{A_k \in A} tf(A_k;d)$$

Comparing this variation to the full model PW, helps us to understand the effect of using proximity information for weighting the pronoun occurrence as well as the effect of weighting pronoun sets based on their relevance to the query. In fact, we aim to see how the performance changes when we turn off these two weighting features and just add the pronoun frequencies to the query term frequency.

Variation W

Variation W also uses standard term frequency to calculate $tf_{anaphoric}(w;d)$. However, it uses $p(A_k|Q)$ to weight the pronouns sets' frequency:

$$tf_{anaphoric}(w;d) = \sum_{A_k \in A} P(A_k|Q)tf(A_k;d),$$

In this variation, we just turn off the proximity weighting but keep the pronoun weighting component. Here, for every pronoun set, we count the occurrences of all pronouns in that set, even when they are far away from any query term, however, the frequency of every pronoun (e. g. she) is weighted by

the probability of its associated pronoun set given the query (e. g. $P(A_f|Q)$). Section 3.1 explains how to estimate this probability.

Variation P

In variation P, we turn off the pronoun type weighting and investigate the effect of using the proximity information to estimate the probability that a pronoun is coreferential with the query term. We therefore estimate the pronoun count as follows:

$$tf_{anaphoric}(w;d) = \sum_{A_k \in A} atf_w(A_k;d),$$

In all variations we consider a ranking of pronoun sets based on the $P(A_k|Q)$. In Table 2, we report the result of different variations of the model. For each variation, we report the result of using all pronoun sets with subscript 3 (i. e. T_3 or W_2, P_2 and PW_2). The result of using only the most probable pronoun set is also reported using subscript 1(i. e. T_1 or W_1, P_1 and PW_1)

Table 2 shows that the best results are achieved using the full setting and all pronoun sets (PW_3). It is interesting that in setting PW, using all pronouns does not degrade the performance compared to using just the highest relevant set of pronouns. The reason is that the pronoun-set weighting of the model along with the proximity method help the system in using more than one set of pronouns for ambiguous queries without adding noise for the non-ambiguous ones (for non-ambiguous queries, just one pronoun sets has a weight close to 1 while the rest are close to zero).

Analysis of different components is conducted by trying variations T, W and P. As we can see in the results of variation T, using all counts of the selected pronoun sets without weighting by proximity information or by feasibility probability leads to statistically significant decreases in the performance. The same result is obtained in variation W, where we weight pronoun-set counts by their feasibility probability but we ignore the proximity information. This way, we may count a pronoun which occurs so far from the query and so may not be referring to the query as a query occurrence. However, compared to T, in W, we ignore or weight less pronouns that are not feasible for the query and this leads to improvement in setting W compared to setting T. We also try the effect of using proximity information without using the feasibility weight, in setting P. We aim to see if the proximity information is enough for capturing the relevance of a pronoun to the query without bothering to weight the pronoun sets. This setting may lead to count a pronoun of *male* category for a *female* query entity, if it occurs in close proximity to the query. As we can see from the result, using just proximity information in weighting the pronoun counts leads to statistically significant improvements

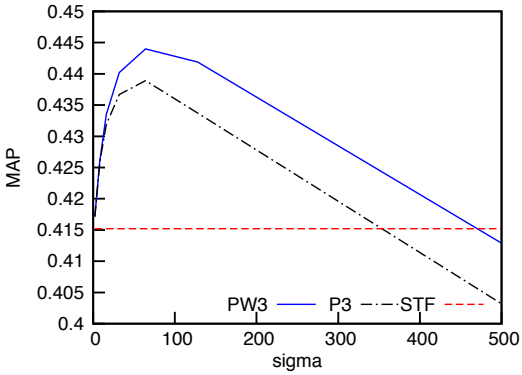


Fig. 2 Sensitivity to the propagation range σ

compared to STF method which ignores the pronoun counts. Although the proximity information alone is enough for improving the performance, adding the feasibility weighting component leads to higher performance. Comparing the performance of P_i with PW_i reveals the importance of considering the feasibility weights specially in case of using more pronoun sets. In fact, when using the most feasible pronoun set (P_1), adding feasibility weight (PW_1) improves the performance, but the improvement is not a lot. However, when we consider two or more set of pronouns, considering the feasibility weight is more effective (e. g. compare P_i with PW_i where $i > 1$).

The proposed proximity-based models have a σ parameter (equation 3) which reflects how far a pronoun and query term can be and still be corefrential. Figure 2 shows the sensitivity (in terms of MAP) of PW_3 and P_3 settings to the different values of σ parameter. We can see that PW_3 has higher MAP and is more stable across different values of σ.

4.3.1 Comparison with Related Work

To the best of our knowledge, the only previous work that considered the counting of pronouns in the query occurrence statistics is the CEEF model in [4]. CEEF is based on the query-entity phrase frequency rather than individual query term frequencies and involves manual identification of part of the query which reflects the query-entity. Therefore the direct comparison of our model with CEEF is not reasonable. Authors in [4] report the MAP value of 0.4139 in case of using CEEF in comparison with ignoring the pronoun counts and using the raw entity frequency which leads to MAP value of 0.3929. Our method PW_3 with MAP 0.4402 also improve over the standard query term frequency with MAP 0.4152. This indicates the advantage of considering pronoun counts even when working at term level instead of entity level. Authors in [4] report further improvement by combing the ranking

of a standard term based retrieval model such as language model with the entity based model. Our model is not comparable with that variation since our model is not based on entity phrase and there is no point in combining it with another term level model.

5 Conclusion and Future Work

In this paper we proposed a proximity-based method for estimating the count of pronouns that can be coreference with the query terms. We incorporated this anaphoric count in the frequency of query terms and used it for retrieval. The proposed model was shown to be effective over the large and standard BLOG06 collection. For future work we plan to investigate the effect of using more complete sets of anaphoric expressions which goes beyond pronouns. For instance, *documentary* or *movie* may be used to refer to a movie instead of repeating its name. We also plan to apply the same method over the entity frequency instead of individual term frequencies and compare the results with the previous work [4].

References

1. Bishop CM (2007) *Pattern Recognition and Machine Learning (Information Science and Statistics)* Springer, Berlin Heidelberg New York
2. Gerani S, Carman MJ, Crestani F (2010) Proximity-based opinion retrieval. *Proc. of SIGIR'10*, 403–410
3. Macdonald C, Ounis I (2006) *The TREC blogs06 collection: Creating and analysing a blog test collection.* DCS Technical Report Series `http://www.dcs.gla.ac.uk/~craigm/publications/macdonald06creating.pdf`
4. Na SH, Ng HT (2009) A 2-poisson model for probabilistic coreference of named entities for improved text retrieval. *Proc. of SIGIR'09*, 275–282
5. Nam SH, Na SH, Lee Y, Lee JH (2009) Diffpost: Filtering non-relevant content based on content difference between two consecutive blog posts. *Proc. of ECIR'09*, 791–795
6. Robertson S, Walker S, Jones S, Hancock M, Gatford M (1994) Okapi at TREC-3. *Overview of the 3rd Text REtrieval Conference (TREC 3)*, 109–126
7. Santos RL, He B, Macdonald C, Ounis I (2009) Integrating proximity to subjective sentences for blog opinion retrieval. *Proc. of ECIR'09*, 5478:325–336
8. Soon WM, Ng HT, Lim DCY (2001) A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.* 27:521–544
9. Versley Y, Ponzetto S, Poesio M, Eidelman V, Jern A, Smith J, Yang X, Moschitti A (2008) Bart: A modular toolkit for coreference resolution. *Proc. 6th Int. Conf. on Language Resources and Evaluation (LREC'08).* European Language Resources Association (ELRA), Marrakech
10. Zhang W, Yu C, Meng W (2007) Opinion retrieval from blogs. *Proc. of CIKM'07*, 831–840

Derivation of Linguistic Summaries is Inherently Difficult: Can Association Rule Mining Help?

Janusz Kacprzyk[1] and Sławomir Zadrożny[1,2]

Abstract We present first the essence of fuzzy linguistic summaries, indicate their relation to fuzzy queries with linguistic quantifiers, and show a taxonomy of protoforms of linguistic summaries indicating that a general protoform, which corresponds to some type of an IF-THEN rule, parallels the structure and form of an association rule. We show that the use of our fuzzy querying interface makes it possible to operationalize the process of definition, updating and processing of fuzzy terms in linguistic data summaries (fuzzy values, fuzzy relations, fuzzy linguistic quantifiers, etc.) and their corresponding fuzzy association rules of a special type. We develop for them a mining algorithm based on AprioriTID. This is clearly a step towards an effective and efficient method for the generation of linguistic data summaries which is badly needed for their proliferation in practice.

1 Introduction

We deal with *linguistic data(base) summaries* which try to grasp the very meaning of a (usually huge) set of (numeric, in our case) data via a simple and short statement in natural language. For instance, for a (possibly huge) personnel database a short yet informative linguistic summary may be "most young and highly qualified people have high salaries." The need for data summarization, one of the basic capabilities to be possessed by any intelligent system, is a direct result of an abundance of data that is beyond human cognition and comprehension, and that for a human being the only fully natural means of communication is natural language.

[1] Systems Research Institute, Polish Academy of Sciences ul. Newelska 6, 01-447 Warsaw, Poland, {kacprzyk, zadrozny}@ibspan.waw.pl
[2] Technical University of Radom, ul. Malczewskiego 29, 26-600 Radom, Poland

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 291–303.
springerlink.com

We consider the use of linguistic data(base) summaries introduced by Yager [24, 26], then advanced by Kacprzyk and Yager [8], and Kacprzyk, Yager and Zadrożny [9], and implemented in Kacprzyk and Zadrożny [12, 14, 16, 17], which are linguistically quantified propositions. It is worth to mention some other approaches to the linguistic summarization of databases, cf. Bosc *et al.* [4], Dubois and Prade [5], Raschia and Mouaddib [22] or Rasmussen and Yager [23].

Though the concept of linguistic summaries is simple and intuitively appealing, their derivation (mining) is difficult since the linguistic summaries make sense for large sets of data, and may involve a considerable number of linguistic values. Moreover, at a conceptual level, an automatic expression of the real human interest and intention with respect to a linguistic summary is questionable. We will adopt our general approach (cf. Kacprzyk and Zadrożny [12, 13]) of an interactive approach via the use of our FQUERY for Access, a fuzzy querying add-on (see Kacprzyk and Zadrożny's [10, 11, 15] and also Zadrożny *et al.* [31]).

First, we show that by relating various types of linguistic summaries to fuzzy queries, with various known and sought elements, we end up with a hierarchy of Zadeh's [29] protoforms of linguistic data summaries. We mention some ways for an automatic generation of linguistic summaries for various protoforms, and indicate that the most general one closely resembles some special types of IF-THEN rules. We show that the mining of such general protoforms of linguistic data summaries can be implemented via the mining of traditional association rules for which many quite well known and powerful algorithms are known.

2 Linguistic Summaries via Fuzzy Logic with Linguistic Quantifiers

In the basic Yager's [24] approach, in its constructive form by Kacprzyk and Yager [8], and Kacprzyk, Yager and Zadrożny [9], and implemented in Kacprzyk and Zadrożny [12, 15, 16], we have: (1) V, a quality (attribute) of interest, e.g. salary in a database of workers, (2) a set of objects (records) y_i that manifest quality V, e.g. the set of workers; hence $V(y_i)$ are values of quality V for objects y_i, and (3) $Y = \{V(y_1), \ldots, V(y_m)\}$ is a set of m pieces of data (the "database" in question). A linguistic summary of a data set consists of:

- a *summarizer* S (e.g. young, extendable to young and well paid, etc.),
- a *quantity in agreement* Q given as a fuzzy linguistic quantifier (e.g. *most*),
- *truth degree* T — e.g. 0.7, as, e.g., "T(*most employees are young*) = 0.7".

The calculation of the truth degree is equivalent to the calculation of the truth value (from $[0, 1]$) of a linguistically quantified statement which may be done

by using Zadeh's [28] calculus of linguistically quantified propositions (cf. Zadeh and Kacprzyk [30]) which will be used here; cf. also Yager's [25] OWA operators (cf. also Yager and Kacprzyk [27]). Zadeh's calculus of linguistically quantified propositions makes it possible to calculate the truth value of the propositions:

$$Q_{y\in Y}\, S(y) \quad \text{(e.g. "Most elements of } Y \text{ possess property } S\text{")} \tag{1}$$

$$Q_{y\in Y}\, K\, S(y) \quad \text{(e.g. "Most elements of } Y \text{ with property } K \text{ possess also property } S\text{")} \tag{2}$$

using the following formulas, respectively:

$$\text{truth}(Q\, S(y)) = \mu_Q\left(\frac{\sum \text{Count}(S)}{\sum \text{Count}(Y)}\right) = \mu_Q\left(\frac{1}{m}\sum_{i=1}^{m}\mu_S(y_i)\right) \tag{3}$$

$$\text{truth}(Q\, K\, S(y)) = \mu_Q\left(\frac{\sum \text{Count}(S\cap K)}{\sum \text{Count}(K)}\right) = \mu_Q\left(\frac{\sum_{i=1}^{m}\mu_S(y_i)\wedge\mu_k(y_i)}{\sum_{i=1}^{m}\mu_K(y_i)}\right) \tag{4}$$

where $m = \text{card}(Y)$, $\sum \text{Count}(A) = \sum_{y_i\in Y}\mu_A(y_i)$, $\sum_{i=1}^{m}\mu_k(y_i) \neq 0$, and $\wedge$ is a t-norm.

Formula (2) represents a richer form of a linguistic summary which covers a fuzzy subset of a "database" Y, defined by a fuzzy predicate K, a *qualifier*.

The basic validity criterion, i.e. the truth degree T, calculated using (3) or (4) is certainly the most important but does not grasp all aspects of a linguistic summary. As to some other quality (validity) criteria, e,g., Yager [24] proposed a measure of informativeness, and then five additional measures have been proposed by Kacprzyk and Yager [8] and Kacprzyk, Yager and Zadrożny [9]: truth, degrees of imprecision, covering and appropriateness, and a length of a summary. For more measures, see Kacprzyk, Wilbik and Zadrożny [7]. The problem is how to generate the best summary (or summaries). An exhaustive search can obviously be computationally prohibitive, and some implicit enumeration type schemes should be used to be dealt with in the next section.

3 A Natural Relation between Linguistic Summaries and Fuzzy Querying – A Protoform based Analysis

Since it is difficult to automatically detect what (in the sense of a linguistic summary) is interesting, intended, useful, etc. to the user, Kacprzyk and Zadrożny [13] proposed an interactive approach for the definition of elements of an intended linguistic summary via a user interface of a fuzzy querying add-on. The roots are our previous papers on the use of fuzzy logic in querying databases (cf. Kacprzyk and Ziółkowski [19], Kacprzyk, Zadrożny and

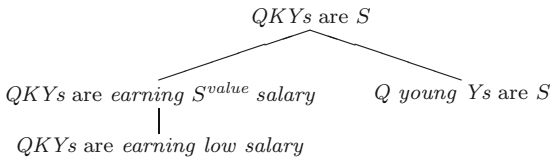


Fig. 1 An example of a part of a hierarchy of protoforms

Ziółkowski [18]) via imprecise requests which led to our FQUERY for Access, an add-in to Microsoft Access® that makes it possible to use fuzzy linguistic terms in database queries such as *numerical fuzzy values*, exemplified by *low* in "profitability is *low*", *fuzzy relations*, exemplified by *much greater than* in "income is *much greater than* spending", and *linguistic quantifiers*, exemplified by *most* in "*most* conditions have to be met".

These fuzzy linguistic terms are building blocks of fuzzy queries in our approach and are represented as fuzzy sets. Notably, *linguistic quantifiers* provide for a more flexible aggregation of simple conditions in queries. For example, instead of requiring that all simple conditions are met, one may indicate that *most* of them are to be met. Clearly, linguistic terms have to be defined and stored internally. This was implemented in our FQUERY for Access package, an add-in to Microsoft Access (cf. Kacprzyk and Zadrożny [10, 11, 15]).

Obviously, fuzzy queries directly correspond to linguistic summaries. Thus, the derivation of a linguistic summary may proceed as: (1) the user formulates a set of linguistic summaries of interest (relevance) using the fuzzy querying add-on, (2) the system retrieves records from the database and calculates the validity of each summary adopted, and (3) a most appropriate linguistic summary is chosen. Operationally, to derive linguistic summaries, some standardized forms of linguistic summaries would be desirable, and this is provided by Zadeh's *protoform* viewed as an abstract prototype of a linguistic summary given by (1) or (2).

For the generation of linguistic summaries it is convenient to consider the summarizer (and the qualifier) as an abstract fuzzy logic statement "X IS A", where X is a placeholder for an attribute of objects in Y and A is a placeholder for a fuzzy set (linguistic term) determining its value as, e.g., "age IS young" or also "salary IS A". Two former summarizers are fully instantiated, while the latter still contains an abstract form of the attribute value (A).

Since the protoforms may form a hierarchy, we can define lower level (less abstract) protoforms, for instance replacing Q by a specific linguistic quantifier, "most", and we get: "Most Ys are S" for (1) and "Most KYs are S" for (2). Zadeh's protoforms may conveniently be used as a fundamental element of the user interface in that the user selects a protoform of a linguistic summary from that hierarchy and then the system instantiates the selected protoform in all possible ways, replacing abstract symbols by chosen fuzzy values and linguistic quantifiers stored in a dictionary. A part of such a hierarchy of protoforms is shown in Figure 1. At the top we have a completely

Table 1 A taxonomy of linguistic summaries

Type	Given	Sought	Remarks
1	S	Q	Simple summaries through ad-hoc queries
2	$S\ K$	Q	Conditional summaries through ad-hoc queries
3	$Q\ S^{structure}$	S^{value}	Simple value oriented summaries
4	$Q\ S^{structure}\ B$	S^{value}	Conditional value oriented summaries
5	Nothing	$S\ K\ Q$	General fuzzy rules

abstract protoform; in a protoform to the right, the qualifier K is instantiated to "age IS young"; in the one to the left, summarizer S is first instantiated to "salary IS S^{value}", i.e., the attribute of the summarizer is selected to be "salary" but its value is not determined; then this protoform is further instantiated to fully specify the summarizer using "low" as the value of "salary".

Thus, the more abstract forms of protoforms correspond to cases in which we assume less about the summaries sought. At the one extreme, we: (1) assume a totally abstract top protoform, or (2) assume that all elements of a protoform are given, i.e., all attributes and all linguistic terms expressing their values are fixed. In case 1 data summarization by a "brute force" full search would be extremely time-consuming, but might produce interesting, unexpected views on data, and in case 2 the user is in fact guessing a good candidate summary but the evaluation is simple, by answering a (fuzzy) query; this is related to ad hoc queries.

This classification may be shown as in Table 1 in which 5 basic types of linguistic summaries are shown, corresponding to protoforms of a more and more abstract form; $S^{structure}$ denotes that attributes and their connection in a summary are known, while S^{value} denotes the values of the attributes sought.

Type 1 summaries may be easily obtained by a simple extension of fuzzy querying. The user has to construct a query, a candidate summary, and it has to be determined what is the fraction of rows matching this query and what linguistic quantifier best denotes this fraction. A Type 2 summary is a straightforward extension of Type 1. Type 3 summaries require much more effort as their primary goal is to determine typical or exceptional, depending on the quantifier, values of an attribute. A Type 4 summary is meant to find typical (exceptional) values for some, possibly fuzzy, subset of rows. Computationally, Type 5 summaries represent the most general form considered here: fuzzy rules describing dependencies between specific values of particular attributes. Type 1 and 3 summaries have been implemented as an extension to Kacprzyk and Zadrożny's [12] FQUERY for Access. Two approaches to Type 5 summaries generation have been proposed. First, a subset of such summaries may be obtained by analogy with association rules concept and employing their efficient algorithms. Second, genetic algorithms may be used to search the space of summaries (cf. George and Srikant [6]. In the next

section we discuss in a more detailed way one more special case of Type 5 summary, for which computationally efficient algorithms are known.

4 Linguistic Summaries and Association Rules: An Intrinsic Relationship

In this section, we will follow another path by investigating the similarity of our Type 5 summaries to association rules [1]. Originally, the association rules were defined for transactional data and binary valued attributes in the following form:

$$A_1 \wedge A_2 \wedge \ldots \wedge A_n \rightarrow A_{n+1} \tag{5}$$

An association rule states that if in a database row all the attributes from $\{A_1, A_2, \ldots, A_n\}$ take on value 1, then also attribute A_{n+1} is expected to take on value 1.

A row in a database (table) is said to *support* a set of attributes $\{A_i\}_{i \in I}$ if all attributes from the set take on in this row value 1. There are two main quality measures for the association rule (5): the *support* which is the fraction of the number of rows supporting the set of attributes $\{A_i\}$, $i \in \{1, \ldots, n+1\}$, in a database (table), and the *confidence* which is the fraction of the rows supporting $\{A_i\}$, $i \in \{1, \ldots, n+1\}$ among all rows supporting $\{A_i\}$, $i \in \{1, \ldots, n\}$. While the support determines a statistical significance of a rule, the confidence measures its strength. Usually, we are interested in rules having values of the support above some minimal threshold and a high value of the confidence. Many algorithms for finding all association rules possessing a required support measure were devised, see, e.g. Agrawal and Srikant [1], Borgelt and Kruse [3].

As to most obvious extensions of the initial form of the association rule, one can mention the following ones: (1) the right-hand side, like the left-hand side, may contain a conjunction of the attributes instead of just one attribute, (2) many-valued scalar values and their hierarchies may be used, (3) numerical, real-valued attributes may be used leading to the *quantitative association rules*, and (4) some constraints may be imposed on combinations of attributes in rules. In view of (1) and (3) the initial scheme of an association rule may be rewritten as:

$$A_1 = a_1 \wedge A_2 = a_2 \wedge \ldots \wedge A_n = a_n \rightarrow A_{n+1} \wedge \ldots \wedge A_{n+m} = a_{n+m} \tag{6}$$

Clearly, the association rules may be interpreted as a special case of the linguistic summaries. Namely, the antecedent and consequent of (5) correspond to qualifier K and summarizer S of (2), respectively. The confidence of a rule is related to the combination of the linguistic quantifier and truth degree of (2). The general form a linguistic summary assumes a summarizer S to be a formula, atomic or complex. It is easy to see that the structure of the quali-

fier and the summarizer available in the case of association rules is somehow limited but this simplicity should increase a chance of existence of efficient algorithms for rule generation.

In our previous work (cf. Kacprzyk and Zadrożny [14]) we implemented the mining of linguistic summaries corresponding to the association rule (6) within the framework of our fuzzy querying package FQUERY for Access. For that purpose we generalized (6) to:

$$A_1 \text{ IS } f_1 \wedge \ldots \wedge A_n \text{ IS } f_n \rightarrow A_{n+1} \text{ IS } f_{n+1} \wedge \ldots \wedge A_{n+m} \text{ IS } f_{n+m} \tag{7}$$

This boils down to the use of *fuzzy values* f_i instead of crisp values which clearly implies that a *fuzzy association rule* is obtained.

Here we propose two extensions to this form of *fuzzy association rule*. First, we enrich the structure of an atomic condition: A_i IS f_i, (an *item* meant in the terminology of the association rules), by allowing it to take the following form:

$$A_i \text{ IS } (f_{j1} \vee \ldots \vee f_{jk}) \tag{8}$$

where f_{jl} are some fuzzy values defined over the domain of the attribute A_i. Thus, we propose to make it possible to use a range of fuzzy values which corresponds to the case of quantitative crisp association rule (cf., for instance, [21]). The second extension of (6) is to use a flexible aggregation operator in the summarizer and/or qualifier formula. This leads to the following form of the atomic condition:

$$Q \text{ of } (A_1 \text{ IS } f_1, A_2 \text{ IS } f_2, \ldots, A_n \text{ IS } f_n) \tag{9}$$

In practical cases this will be most often the only atomic formula in an antecedent or consequent of an association rule. However, it may be just a part of a compound formula.

Such a simple, yet intuitively appealing and highly constructive extension of the fuzzy association rules is to a large extent implied by the capabilities of our FQUERY for Access. This will be briefly presented in the next section.

5 Mining Fuzzy Association Rules by using a Fuzzy Querying Interface

Now, we present how FQUERY for Access can be used to find the fuzzy association rules corresponding to the Type 5 protoforms of linguistic summaries.

Basically, FQUERY for Access supports various linguistic terms, as mentioned: *numerical fuzzy values* ("*low*"), *scalar fuzzy values* ("*Central Europe*"), *fuzzy relations* ("*much greater than*"), and *linguistic quantifiers*

("*most*") handled using Zadeh's [28] calculus of linguistically quantified propositions.

First, notice that our fuzzy querying interface offers a practical solution to some problems faced by the "classical" [21] search for the association. Namely, the quantitative association rules usually require discretization of the attributes obtained via a partition of its domain into a number of intervals. Then, each interval is treated as an additional binary attribute and then many known algorithms for the generation of classical association rules may be employed. It is not obvious how to do such a partition but having a domain covered by a number of overlapping fuzzy values (defined and tested for the purposes of fuzzy querying on many former cases of applications) the partition is readily available.

The combination of fuzzy querying and mining within the same interface seems to yield a synergetic effect, and from the viewpoint of software engineering it is advantageous to employ some of the modules to support both functions. The user interface is the same as for fuzzy querying. However, for computational efficiency reasons, the very algorithm of fuzzy association rules mining is implemented as a separate executable.

Our implementation of association rules is based on the Agrawal and Srikant's AprioriTID algorithm [1] (cf. Borgelt and Kruse [3]) which works in two steps: first it finds *frequent itemsets* and then produces rules from each itemset. The second step is relatively easy, hence we will focus on the first one.

An itemset is a conjunction of the items of the form (8) or (9). A row in the database (table) is said to *support* an itemset if the corresponding conjunction "is true" (the degree of satisfaction exceed some threshold) for this row. An itemset containing k items is called a k-itemset. The algorithm starts with the evaluation of 1-itemsets. These itemsets which are not supported by sufficient number (*minsup*) of rows are deleted. Previously, we assumed 1-itemsets only in the form (7). In order to implement items of the form (8) or (9) we have to extend this step. We treat as the 1-itemsets also the items like A_i IS $(f_{j1} \vee \ldots \vee f_{jk})$ and Q of $(A_1$ IS f_1, A_2 IS $f_2, \ldots, A_n$ IS $f_n)$. More precisely, first, only the "regular" 1-itemsets (7) are counted, i.e., a full scan of the database (table) is done and the frequency of appearance of all items is calculated. Then, the 1-itemsets of type (8) are constructed but only such f_{ij} are taken into account that have the support greater than some value (a parameter of the method, in addition to *minsup* and *minconf*) higher than 0 and less than *minsup*. For example, if a regular 1-itemset "salary IS high" gets a very low support, then we will not construct either "salary IS medium or high" or "salary IS low or high" 1-itemsets. This helps reduce the time and memory complexity of the algorithm.

Such a reduction is even more important in case of the implementation of the 1-itemsets of type (9). Basically, we should take into account all subsets of the regular 1-itemsets and all possible quantifiers Q. This would be computationally intractable and in fact require a kind of a recursive use of AprioriTID

in the first step. Thus, in our implementation we limit ourselves to just one, fixed quantifier. Moreover, for obvious reasons, we take into account only such subsets of regular items that: all refer to different attributes, and there is enough number of them to make quantification meaningful. Thus, we will, e.g., neither construct a 1-itemset of the form "*most* (*salary* IS *high, salary* IS *low*, ...)" nor "*most* (*salary* IS *high*, *age* IS *high*).

Having the 1-itemsets of type (8) constructed we calculate their support. We assume that no row supports two different fuzzy values for the same attribute. Then, the support for A_i IS $(f_{j1} \vee \ldots \vee f_{jk})$ is just the sum of supports for A_i IS f_{jl}, $l = 1, \ldots, k$; that we calculated earlier. Now the 1-itemsets of type (9) are constructed. They may use both the regular 1-itemsets as well as the 1-itemsets of type (8), e.g., "Most (A_1 IS $(f_{11} \vee \ldots \vee f_{1k})$, A_2 IS f_2, ...)" are allowed. The support for the 1-itemsets of type (9) is then calculated. However, due to the use of AprioriTID another full scan of the database is not needed. During the first scan we have recorded in some data structures the IDs of rows supporting the particular regular 1-itemsets and now it is enough to operate on these structures. Obviously, it does increase memory complexity of the algorithm. Then, the algorithm proceeds as usual [1] generating and evaluating the k-itemsets for $k = 2, 3, \ldots$. The only additional effort is needed to guarantee that no itemset produced twice refers to the same attribute, e.g., the 2-itemset "*salary* IS *high* AND *salary* IS *medium*" has to be excluded. Finally, all frequent itemsets found are taken into account when producing association rules of the confidence at least equal to the required value (*minconf*).

We deal with the real valued attributes so that for each such an attribute and each fuzzy value defined for it we introduce a new items which may be treated as binary, i.e., appearing in a row or not. In this respect, practically only a limited number of fuzzy values per attribute (say 3) leads to computationally tractable mining tasks.

The implementation of the algorithm for the mining of linguistic summaries via extended fuzzy association rules may be presented as follows:

Step 1: *Selection of the attributes and fuzzy values*
The user chooses the attributes to be used, i.e. builds a query referring to the attributes to be taken into account. Then, the user initiates the data summarization process, sets the parameters (*minsup*, *minconf*, minimum support, ...) and the system automatically performs the rest of the steps.

Step 2: *Construction of the items*
For each pair – of the selected attributes and fuzzy values — the system creates an item, as described earlier.

Step 3: *Forming the data set and starting external application for fuzzy linguistic rules mining*
The items constructed are numbered. Then, the data set is produced describing each row with numbers of items supported by it. The calculations proceed by the fuzzy querying module. When the data

set is ready, an external application is started with this data set given on input.

Step 4: *Calculation of the support for the regular 1-itemsets*
An external application reads the input data set and immediately calculates support for the regular 1-itemsets. It also records for each 1-itemset numbers (IDs) of rows supporting it.

Step 5: *Construction of the 1-itemsets of type (8) and calculation of their support*
Only the regular 1-itemsets of the support higher than a user-specified threshold are taken into account. The number of 1-itemsets of this type produced for a given attribute depends on the number of fuzzy values defined for it. The support is obtained by summing up the support of the constituent regular 1-itemsets. All new 1-itemsets are numbered.

Step 6: *Pruning of the set of 1-itemsets*
All itemsets with the support lower than the support threshold (*minsup*) are discarded. Additionally, also itemsets with the support higher than another threshold, an item omit threshold, are discarded since the items present in almost all records contribute nothing interesting to the rules produced.

Step 7: *Construction of the 1-itemsets of type (9), calculation of their support and pruning*
Both the regular and 1-itemsets added in Step 5 are considered; we refer to them jointly as simple 1-itemsets. The 1-itemsets constructed are identified with lists of the constituent simple 1-itemsets. The lists are ordered lexicographically which makes the process of generation more efficient. The support is computed for the itemsets generated and those below the minsup threshold are discarded.
All itemsets produced so far and passing the pruning constitute the collection of 1-itemsets.
SET $k = 2$

Step 8: *Generate the k-itemsets*
They are generated from the frequent $(k-1)$-itemsets as in AprioriTID. Pairs of the frequent $(k-1)$-itemsets of the form $A_1 \wedge A_2 \wedge \ldots \wedge A_{k-1}$ and $B_1 \wedge B_2 \wedge \ldots \wedge B_{k-1}$, where $A_i = B_i$ for $i = 1, \ldots, k-2$, are sought. Then, a new k-itemset of the form $A_1 \wedge A_2 \wedge \ldots \wedge A_{k-1} \wedge B_{k-1}$ is generated. In the original algorithm, the rules generated in such a way are additionally tested and possibly eliminated before Step 7. On the other hand, we add another k-itemset generation limitation, namely the items A_{k-1} and B_{k-1} have to correspond to different original attributes. This is obvious if the items A_{k-1} and B_{k-1} are regular. Otherwise, by identifying an item of type (8) or (9) with a list (set) of attributes referred to within it, we require the intersection of these sets to be empty.

Step 9: *Calculate the support for all the k-itemsets*
The calculation is based on the recorded numbers (IDs) of rows supporting the particular $(k-1)$-itemsets. The similar data on the supporting rows is produced for the k-itemsets.

Step 10: *Pruning of the set of k-itemsets (as in Step 6)*
As a result we obtain the frequent k-itemsets.
IF the set of k-itemsets is void THEN GOTO Step 11.
SET $k = k + 1$; GOTO Step 8.

Step 11: *Generate rules from the frequent l-itemsets,* $l = 1, \ldots, k-1$ *and output them to the file which is done by an external application*

Step 12: *Display the results*

The number of the rules produced is usually huge. Some counter-measures have to be undertaken as, e.g., approaches to the representation of concise association rules [20]. We adopt a simple yet efficient pruning scheme. A rule R_1 is pruned if there exists another rule R_2 such that the 3 conditions are met simultaneously: (1) the antecedent of R_2 is a subset of that of R_1, (2) the consequent of R_1 is a subset of that of R_2, (3) the confidence of R_2 is not less than that of R_1. This leads to asubstantial, lossless reduction of the number of rules.

6 Remarks on an Implementation

Elements of the method for generating linguistic data summaries through association rule mining have been used by us in a number of applications, and we have employed therein the AprioriTID in Borgelt's implementation (cf. `http://www.borgelt.net/apriori.html`). One of the most intuitively appealing examples was related to the use of linguistic data summaries to a human consistent analysis of data related to the innovativeness of Polish companies (cf. Baczko, Kacprzyk and Zadrożny [2]). The values of each attribute were described by three linguistic terms: low, medium and high, and if needed, the original values of selected numerical attributes were replaced by their best matching linguistic terms. The definition of linguistic terms was supported by FQUERY for Access. The linguistic quantifier "most" was used in the generated summaries. The set of transformed data was processed by AprioriTID in Borgelt's implementation. We obtained a lot of very interesting linguistic summaries exemplified by: "Most companies having high net revenues from sales and equivalent in 2004 had high total assets in 2004", "Most of the companies having at least a few points (scores) for their RTD related activities in 2006 had also some points for that in 2005", "Most companies having some points related to patents registered in 2006 AND some points for their RTD related activities in 2005 had also some points for RTD related activities in 2006", etc. Thus, in general, companies being active in the RTD field in 2005 did not necessarily continue to do so in 2006. However,

those with some patents in 2006 usually also had RTD related activities in 2006".

7 Concluding Remarks

The use of our fuzzy querying interface has made it possible to operationalize the process of definition, updating and processing of fuzzy terms which are meant to exist in linguistic data summaries (fuzzy values, fuzzy, relations, fuzzy linguistic quantifiers, etc.) and, if we limit our attention to linguistic summaries following a special, very general protoform, which are meant to be in fuzzy association rules which correspond to those special linguistic summaries. And, for those association rules we have developed a mining algorithm based on the idea of AprioriTID. This is clearly a step towards an effective and efficient method for the generation of linguistic data summaries which is badly needed for their proliferation in practice. Both the use of association rule mining, in which much software and experience exists, and our practical experience outlined clearly indicate that this approach to the derivation of linguistic summaries may be effective and efficient.

References

1. Agrawal R, Srikant R (1994) Fast Algorithms for Mining Association Rules. *Proc. 20th Int. Conf. on Very Large Databases (VLDB 1994, Santiago de Chile)*, 487-499. Morgan Kaufmann, San Mateo, CA, USA
2. Baczko T, Kacprzyk J, Zadrożny S (2011) Towards Knowledge Driven Individual Integrated Indicators of Innovativeness. In: Jozefczyk J, Orski D (eds.) *Knowledge-Based Intelligent System Advancements: Systemic and Cybernetic Approaches*, 129–140. IGI Global, Hershey, USA
3. Borgelt C, Kruse R (2002) Induction of Association Rules: Apriori Implementation. *Proc. 15th Conf. on Comp. Statistics (Compstat 2002, Berlin, Germany)*, 395-400. Physica Verlag, Heidelberg, Germany
4. Bosc P, Dubois D, Pivert O, Prade H, de Calmes M (2002) Fuzzy Summarization of Data Using Fuzzy Cardinalities, 1553-1559. *Proc. IPMU'2002 (Annecy, France)*
5. Dubois D, Prade H (1992) Gradual Rules in Approximate Reasoning. *Information Sciences* 61:103-122
6. George R, Srikanth R (1996) Data Summarization Using Genetic Algorithms and Fuzzy Logic. In: Herrera F, Verdegay JL (eds.) *Genetic Algorithms and Soft Computing*, 599-611. Physica-Verlag, Heidelberg, Germany
7. Kacprzyk J, Wilbik A, Zadrożny S (2008) Linguistic Summarization of Time Series Using a Fuzzy Quantifier Driven Aggregation. *Fuzzy Sets and Systems* 159:1485-1499
8. Kacprzyk J, Yager RR (2001) Linguistic Summaries of Data Using Fuzzy Logic. *Int. Journal of General Systems* 30:133-154
9. Kacprzyk J, Yager RR, Zadrożny S (2000) A Fuzzy Logic Based Approach to Linguistic Summaries of Databases. *Int. Journal of Applied Mathematics and Computer Science* 10:813-834

10. Kacprzyk J, Zadrożny S (1995) Fuzzy queries in Microsoft Access v.2. *Proc. FUZZ-IEEE/IFES '95 (Yokohama, Japan), Workshop on Fuzzy Database Systems and Information Retrieval*, 61-66
11. Kacprzyk J, Zadrożny S (1995) FQUERY for Access: Fuzzy Querying for a Windows-based DBMS. In: Bosc P, Kacprzyk J (eds.) *Fuzziness in Database Management Systems*, 415-433. Physica-Verlag, Heidelberg, Germany
12. Kacprzyk J, Zadrożny S (2000) On Combining Intelligent Querying and Data Mining Using Fuzzy Logic Concepts. In: Bordogna G, Pasi G (eds.) *Recent Research Issues on the Management of Fuzziness in Databases*, 67-81. Physica-Verlag, Heidelberg, Germany
13. Kacprzyk J, Zadrożny S (2001) Data Mining via Linguistic Summaries of Databases: An Interactive Approach. In: Ding L (ed.) *A New Paradigm of Knowledge Engineering by Soft Computing*, 325-345. World Scientific, Singapore
14. Kacprzyk J, Zadrożny S (2001) Fuzzy Linguistic Summaries via Association Rules. In: Kandel A, Last M, Bunke H (eds.) *Data Mining and Computational Intelligence*, 115-139. PhysicaVerlag, Heidelberg, Germany
15. Kacprzyk J, Zadrożny S (2001) CWW in Intelligent Database Querying: Standalone and Internet-based Applications. *Information Sciences* 34:71-109
16. Kacprzyk J, Zadrożny S (2003) Linguistic Summarization of Data Sets Using Association Rules. *Proc. FUZZ-IEEE'03 (St. Louis, USA)*, 702-707.
17. Kacprzyk J, Zadrożny S (2005) Linguistic Database Summaries and Their Protoforms: Towards Natural Language based Knowledge Discovery Tools. *Information Sciences* 173(4):281-304
18. Kacprzyk J, Zadrożny S, Ziółkowski A (1989) FQUERY III+: A 'Human Consistent' Database Querying System based on Fuzzy Logic with Linguistic Quantifiers. *Information Systems* 6:443-453
19. Kacprzyk J, Ziółkowski A (1986) Database Queries with Fuzzy Linguistic Quantifiers. *IEEE Transactions on Systems, Man and Cybernetics* 16:474-479
20. Kryszkiewicz M (2001) Concise Representation of Frequent Patterns based on Disjunction-free Generators. *Proc. of ICDM'2001, San Jose, USA*, 305-312
21. Srikant R, Agrawal R (1996) Mining Quantitative Association Rules in Large Relational Tables. *Proc. of ACM-SIGMOD Conf. on Manag. of Data, Montreal, Canada*
22. Raschia G, Mouaddib N (2002) SAINTETIQ: A Fuzzy Set based Approach to Database Summarization. *Fuzzy Sets and Systems* 129:137-162
23. Rasmussen D, Yager RR (1999) Finding Fuzzy and Gradual Functional Dependencies with summarySQL. *Fuzzy Sets and Systems* 106:131-142
24. Yager RR (1982) A New Approach to the Summarization of Data. *Information Sciences* 28:69-86
25. Yager RR (1988) On Ordered Weighted Averaging Operators in Multicriteria Decision Making. *IEEE Trans. on Systems, Man and Cybern.*, 183-190
26. Yager RR (1996) Database Discovery using Fuzzy Sets. *Int. Journal of Intelligent Systems* 11:691-712
27. Yager RR, Kacprzyk J (1997) *The Ordered Weighted Averaging Operators: Theory and Applications*. Kluwer, Boston, USA
28. Zadeh LA (1983) A Computational Approach to Fuzzy Quantifiers in Natural Languages. *Computers and Maths with Appls.* 9:149-184
29. Zadeh LA (2002) A Prototype-centered Approach to Adding Deduction Capabilities to Search Engines — The Concept of a Protoform. BISC Seminar, Univ. of California, Berkeley, USA
30. Zadeh LA, Kacprzyk J., eds. (1999) *Computing with Words in Information/Intelligent Systems, 1. Foundations. 2. Applications*. Physica-Verlag, Heidelberg/New York, Germany/USA
31. Zadrożny S, De Tré G, De Caluwe R, Kacprzyk J (2008) An Overview of Fuzzy Approaches to Flexible Database Querying. In: Galindo J (ed.) *Handbook of Research on Fuzzy Information Processing in Databases*, 34-53. Idea Group Inc.

Mining Local Connectivity Patterns in fMRI Data

Kristian Loewe[1], Marcus Grueschow[2], and Christian Borgelt[3]

Abstract A core task in the analysis of functional magnetic resonance imaging (fMRI) data is to detect groups of voxels that exhibit synchronous activity while the subject is performing a certain task. Synchronous activity is typically interpreted as functional connectivity between brain regions. We compare classical approaches like statistical parametric mapping (SPM) and some new approaches that are loosely based on frequent pattern mining principles, but restricted to the local neighborhood of a voxel. In particular, we examine how a soft notion of activity (rather than a binary one) can be modeled and exploited in the analysis process. In addition, we explore a fault-tolerant notion of synchronous activity of groups of voxels in both the binary and the soft/fuzzy activity setting. We apply the methods to fMRI data from a visual stimulus experiment to demonstrate their usefulness.

1 Introduction

The localization and analysis of brain activity is a major objective in cognitive neuroscience. Functional magnetic resonance imaging (fMRI) provides an indirect, but non-invasive means to measure brain activity in vivo. Essentially, time series of three-dimensional (3D) brain-images are acquired, in which each volumetric pixel (or *voxel* for short) represents a cuboid of tissue. Inferences about brain activity rest on the following principle: neuronal activity entails the consumption of oxygen and thus the supply of the

[1] Department of Knowledge and Language Processing, University of Magdeburg, D-39106 Magdeburg, Germany, kristian.loewe@gmx.net
[2] Laboratory for Social and Neural Systems Research, Department of Economics, University of Zürich, CH-8006 Zürich, Switzerland, marcus.grueschow@econ.uzh.ch
[3] European Centre for Soft Computing, c/ Gonzalo Gutiérrez Quirós s/n, E-33600 Mieres (Asturias), Spain, christian.borgelt@softcomputing.es

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 305–317.
springerlink.com

relevant area with oxygenated blood. The different magnetic properties of oxygenated blood in comparison to deoxygenated blood result in observable signal changes in the time series of the relevant voxels, which are exploited as an indirect indicator of neuronal activity. This is known as the blood oxygen level dependent (BOLD) effect [9]. For an excellent review of neurovascular coupling and its effect on the BOLD signal see [8].

Typical task-related fMRI experiments are designed and conducted on the grounds of certain hypotheses about brain functions which are subsequently tested using regression-based statistics. To this end, most often general linear models (GLM) based on canonical hemodynamic response functions (cHRF) are fitted to each individual voxel time series in order to obtain statistical maps highlighting brain activity related to experimental conditions. In that regard, several *apriori* assumptions are widely accepted by the neuroimaging community. For example, to facilitate comparison between voxels, GLMs are generated and fitted using the same cHRF for all time series, even though hemodynamic responses differ widely across the brain [1, 5]. As a consequence, such an analysis is limited to the testing of *a priori* hypotheses and frequently makes use of *a priori* assumptions, which—in case they are not met—may constrain the significance of the obtained results.

In contrast to this, data-driven approaches might reveal unexpected patterns that in turn could give rise to new hypotheses, while at the same time *a priori* assumptions are avoided as far as possible. For example, recent studies made use of graph-theory in order to derive network characteristics of the brain from interregional functional connectivity matrices, where functional connectivity means the temporal dependence between brain regions [4]. In both, task-related and resting-state settings (where in a resting state no task and no explicit external stimulus is presented) studies indicated that the brain network is organized in a highly clustered way [3, 13]. Recently, this was exploited to compute locally restricted correlations (based on spatial proximity) in order to rapidly identify potential hub regions in the brain [11].

Building in a similar fashion on the strongly clustered brain organization, we present a new, noise-robust, and purely data-driven method targeting local functional connectivity patterns. The proposed approach is applicable to any type of fMRI data (task-related, resting-state etc.) and allows for time-efficient and model-free generation of meaningful brain maps (without making *a priori* assumptions or presuming hypotheses to test).

2 Notions of Activity

FMRI data are series of periodically acquired 3D intensity images. We denote one such series by $\mathbf{i} = (i_1, i_2, \ldots, i_T)$, where T is the number of points in time at which the intensity images i_k, $k \in \{1, \ldots, T\}$, are recorded. The individual images are organized in a regular 3D voxel grid of size $X \times Y \times Z$.

In order to simplify the processing, the voxel coordinates (x, y, z) can be mapped (in an essentially arbitrary, but fixed fashion) to a linear index v with $1 \leq v \leq V = X \cdot Y \cdot Z$. In this way a data set can be represented by a data matrix $\mathbf{S}^{V \times T} = (s_{v,t})$. By $\mathbf{s}_v = s_{v,*} = (s_{v,1}, s_{v,2}, \ldots, s_{v,T})$, that is, the v-th row of $\mathbf{S}$, we denote the time series of voxel v.

As the data are arbitrarily scaled, a meaningful and comparable notion of activity (magnitude) arises only from a relative interpretation. In the following, we derive a binary and a soft notion of voxel activity by considering at any given time the deviation from its temporal average intensity. The binary notion can be seen as a limiting case of the soft notion.

Binary Notion of Activity. We use a simple binary discretization in order to assign to each value in a time series one of the two qualitative states *active* and *inactive*. Formally, the dichotomized time series d_v of a voxel v is given by $\mathbf{d}_v = (\mathrm{d}(s_{v,1}), \mathrm{d}(s_{v,2}), \ldots, \mathrm{d}(s_{v,T})) \in \{0,1\}^T$ induced by the function $\mathrm{d}(s_{v,t}) = \mathrm{H}(s_{v,t} - \tilde{\mathbf{s}}_v)$, $t \in \{1, \ldots, T\}$, where $\tilde{\mathbf{s}}_v$ denotes the median of the values in $\mathbf{s}_v$ and H is the Heaviside step function, defined as $\mathrm{H}(x) = 0$ if $x < 0$ and $\mathrm{H}(x) = 1$ otherwise. In other words, a voxel is regarded as active at a point in time if the corresponding signal intensity value amounts at least to the median of the respective time series. The median was chosen over the mean because it is less sensitive to outliers.

Note that this very simple scheme is naturally open to many points of criticism. For example, it enforces that a voxel is active half of the time, which is clearly debatable. However, it already leads to useful results and thus we defer finding better discretization schemes to future work.

Note also that the concise binary time series representation $\mathbf{d}_v$ can be exploited in order to speed up subsequent analysis through a highly efficient implementation using bit vectors. However, this advantage comes at the expense of the inevitable loss of information due to the discretization.

Soft Notion of Activity. The above discretization implies an extreme sharpening of the signal: whereas the actual signal rises gradually over time, the discretization enforces a sharp instantaneous signal change once the median is exceeded. Effectively, the signal is transformed into a square-wave signal, thus increasing the contrast at the transition sites.

By replacing the Heaviside step function with a sigmoid function (for instance, a logistic function), we introduce a soft notion of activity, which enables a parameterized sharpening of the signal (thus also limiting the information loss). Formally, we transform the time series according to the linear scaling and logistic activation function

$$f_{\mathrm{act}}^{(\beta)}(s_{v,t}) = \left(1 + \exp\left(-\frac{s_{v,t} - \tilde{\mathbf{s}}_v}{\beta(Q_{0.95}(\mathbf{s}_v) - Q_{0.05}(\mathbf{s}_v))}\right)\right)^{-1},$$

where $Q_{0.05}(\mathbf{s}_v)$ and $Q_{0.95}(\mathbf{s}_v)$ denote the 5% and the 95% quantile, respectively, of the time series $\mathbf{s}_v$. Their difference can be seen as an estimate of the

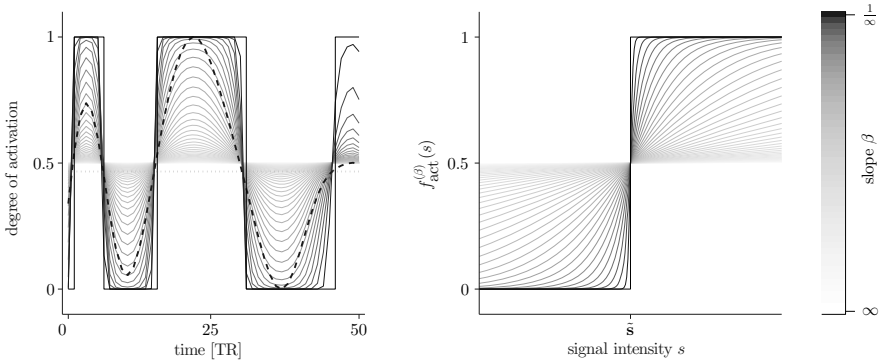


Fig. 1 Application of logistic activation functions differing in their slope β to a time series $\mathbf{s}$. For $\beta \to \frac{1}{\infty}$ this is equivalent to the discretization approach.

range of intensity values, which is more robust than simply using min and max and thus is in line with our choice of the median over the mean.

The strength of the sharpening effect is governed by the slope parameter β. Obviously the binary discretization is obtained as a limiting case of this scheme for $\beta \to \frac{1}{\infty}$. An illustration is shown in Figure 1.

Note that we use the normalization by an estimate of the range of values in order to keep the meaning of the slope parameter independent of the range of values of the time series. Of course, this is also open to criticism, as it removes all information related to the amount of signal change, which may contain valuable information. However, as with the choice of the median as the transition point between inactive and active, we leave further improvements of the activation scheme for future work.

3 Local Connectivity Measures

Recent voxelwise functional connectivity analyses showed a highly clustered organization of the brain in both task-related as well as resting-state settings [3, 13]. Aiming to characterize local connectivity patterns by quantifying the local cohesion strength, we propose new local connectivity measures (LCM) operating on the time series of the enclosing $3 \times 3 \times 3$ cuboid of each voxel. In this scheme each center voxel serves as an identifier of its enclosing cuboid, allowing for the data to be traversed in a sliding 3D window fashion. For this purpose, we denote by

$$N_{26}(v) = \{w \mid v \neq w \wedge \max\{|x_v - x_w|, |y_v - y_w|, |z_v - z_w|\} \leq 1\}$$

the 26 neighbors of a voxel $v = (x_v, y_v, z_v)$ (formed by the 6 voxels sharing a face, the 12 sharing an edge and the 8 sharing a vertex with it). The whole

cuboid of a voxel v is then denoted by $C_{27}(v) = \{v \cup N_{26}(v)\}$ and we only consider voxels with a complete cuboid neighborhood.

A natural approach would be to calculate the average of Pearson's correlation coefficient of all $\frac{27 \cdot 26}{2} = 351$ voxel pairs within an enclosing cuboid. However, each individual correlation—and thus also their mean—might be considerably prone to noise. Especially if there are some voxels that do not participate in the joint activity and thus have low correlations with the other voxels, an existing co-activity pattern may not be discernible.

Therefore, instead of combining pairwise correlations by averaging, we try to obtain a more robust measure by integrating the values of all 27 voxels at each point in time. To be more precise, our idea is to find for each voxel's enclosing cuboid $C_{27}(v)$ those points in time that exhibit synchronous activity of at least a certain (user-specified) number of voxels. Formally, we have

$$\mathrm{LCM}_{\alpha}^{B}(v) = \frac{1}{T} \sum_{t=1}^{T} H\Big(\Big(\sum_{w \in C_{27}(v)} d_{w,t}\Big) - \alpha\Big),$$

where H is the Heaviside step function. "LCM" stands for "local connectivity measure" and the upper index B indicates that it is based on the binary time series $\mathbf{d}_v$. The measure is normalized w.r.t. T, the length of the time series, in order to facilitate comparison between data sets of different length. The parameter $\alpha \in \{1, \ldots, 27\}$ captures the fault-tolerant aspect of this measure: Given a cuboid and a point in time, α active voxels suffice for the corresponding addend to become equal to one (and thus to contribute positively to the co-activity measure).

For functionally independent adjacent voxels we expect to see around 13–14 active voxels at each point in time, because with our discretization scheme (active above and inactive below the median) each voxel is active at half of the points in time and therefore about half of the voxels in a cuboid should be active on average. As a consequence, α should be chosen greater than 14. However, what choice of α is best depends on the level of noise present in the data and the desired contrast between functionally connected and functionally independent 27-cuboids.

Clearly, the number of active voxels can be expected to be significantly higher than 14 at points in time actually showing co-activity. As this co-activity "uses up" some of the active states of the participating voxels, the remaining points in time must possess a lower average number of co-active voxels. In addition, it is plausible to assume that functionally connected voxels also exhibit co-inactivity, that is, possess points in time at which only a relatively low number of voxels are active. This can be exploited to enhance the contrast of the measure by defining

$$\mathrm{LCMd}_{\alpha}^{B}(v) = \mathrm{LCM}_{\alpha}^{B}(v) + (1 - \mathrm{LCM}_{28-\alpha}^{B}(v)).$$

Since $\mathrm{LCM}^B_\alpha(v)$ is the higher, the more co-activity the voxels in a cuboid show, while $(1 - \mathrm{LCM}^B_{28-\alpha}(v))$ is the higher, the more co-inactivity they show, this measure can be expected to be more sensitive.

Using our soft notions of activity, we now define soft analogues of the above measures. In order to handle the activity degrees, we rely on the following reasoning: in a perfect situation, in which all voxels are active at a point in time ($\alpha = 27$), the terms summed over can also be seen as conjunctions of the activity values (active — 1, inactive — 0). In a soft setting this conjunction could be expressed, using a standard fuzzification of conjunctions, by a minimum. The fault-tolerant aspect can then be incorporated by replacing the minimum with a quantile, thus allowing a few activations to be low. This leads to the following measure:

$$\mathrm{LCM}^S_{\alpha,\beta}(v) = \frac{1}{T}\sum_{t=1}^{T} Q_{1-\frac{\alpha-1/2}{27}}\left(\left[f^{(\beta)}_{\mathrm{act}}(s_{w,t}) \;\middle|\; w \in C_{27}(v)\right]\right),$$

where $Q_p([x_1, \ldots, x_k])$ denotes the p-quantile of the data set $[x_1, \ldots, x_{27}]$ (which we do not write as a set in order to allow for multiple voxels having the same activation) such that $1 - \frac{\alpha-1/2}{27}$ selects the α-smallest value. "LCM" again stands for "local connectivity measure" and the upper index S indicates that it is based on a soft notion of activity.

Arguing in the same way as for the binary measure, we can increase the contrast and thus the sensitivity for detecting co-activity by defining

$$\mathrm{LCMd}^S_{\alpha,\beta}(v) = \mathrm{LCM}^S_{\alpha,\beta}(v) + (1 - \mathrm{LCM}^S_{28-\alpha,\beta}(v)).$$

4 Data and Preprocessing

We applied the proposed methods to both artificial data and real fMRI recordings from a task-related experiment.

Artificial Data. In order to analyze the characteristics of the new measures, we generated synthetic data sets of co-active and independent voxels. One sample of a data set consisted of 27 voxel time series corresponding to one $3 \times 3 \times 3$ cuboid of voxels. The co-active samples were created using zero vectors (of length 300) into which blocks of ones were inserted at random locations. The vectors were then convolved with the cHRF included in the software package SPM8[1] for Matlab[2]. Finally, white Gaussian noise (WGN) of given signal-to-noise ratio (SNR) was added. In this way, three data sets of co-active samples were created, corresponding to SNRs of +10, 0 and

[1] SPM8. Wellcome Trust Centre for Neuroimaging, London, UK.
Available at http://www.fil.ion.ucl.ac.uk/spm/.

[2] MATLAB®. The MathWorks Inc., Natick, Massachusetts, USA.

−10 decibel (dB). In addition, we created one data set containing samples of independent voxels using WGN time series.

Real Data. The usefulness of the proposed methods as applied to real data was assessed using task-related fMRI data, acquired using a 7 Tesla MR-scanner (Siemens, Erlangen, Germany) in the context of a study of the visual pathway. Subjects were instructed to focus on a central fixation point, while being exposed to alternating left and right visual hemifield stimulation of different luminance contrasts. Meanwhile, functional data were acquired in volumes of $192 \times 192 \times 27$ voxels at isotropic resolution of 1.1mm edge length using a time resolution of 2s. Details can be found in [12].

The analysis of fMRI data is susceptible to manifold artifacts arising from both physiological and hardware-related sources. It was therefore essential to account for them prior to the actual analysis. Using an online image-reconstruction procedure, all data were motion- and distortion-corrected based on a reference measurement of the local point spread function [14]. As interpolation causes local correlations not originally present in the data, we refrained from spatial smoothing and normalization to a standard brain.

Frequencies below 0.01Hz were removed from the individual voxel time series accounting for low frequency signal intensity drifts caused e.g. by scanner instabilities [10] and physiological artifacts. Non-brain voxels were excluded from further analysis by defining a brain mask using a thresholding procedure based on the means of the voxel time series. For comparative purposes in further analyses the remaining brain voxels were also partitioned into gray matter and non-gray matter voxels by thresholding of a gray matter probability map generated using SPM8 segmentation routines.

GLM Analysis. For comparisons, a conventional GLM analysis of the task-related fMRI data was carried out using SPM8. We applied a statistical model containing boxcar waveforms convolved with a cHRF, representing the left and right visual hemifield stimulation, respectively. Multiple linear regression was then used to generate parameter estimates for each regressor at every voxel. Visual field biased regions in each subject were identified using a contrast of contralateral greater than ipsilateral visual stimulation resulting in a statistical parametric map of t-statistics (SPM_t).

5 Results for Test Data

Three variants of LCM and LCMd were calculated for the four test data sets (see Section 4): LCM^B_α, LCMd^B_α, $\mathrm{LCM}^S_{\alpha,0.05}$, $\mathrm{LCMd}^S_{\alpha,0.05}$, $\mathrm{LCM}^S_{\alpha,0.1}$, and $\mathrm{LCMd}^S_{\alpha,0.1}$ (Figure 2). As anticipated, the LCM is higher for the co-active voxels than for the noise voxels for $\alpha > 14$ while the opposite is true for $\alpha \leq 14$. LCMd exploits both complementary contrasts and thus exhibits an increased sensitivity compared to LCM.

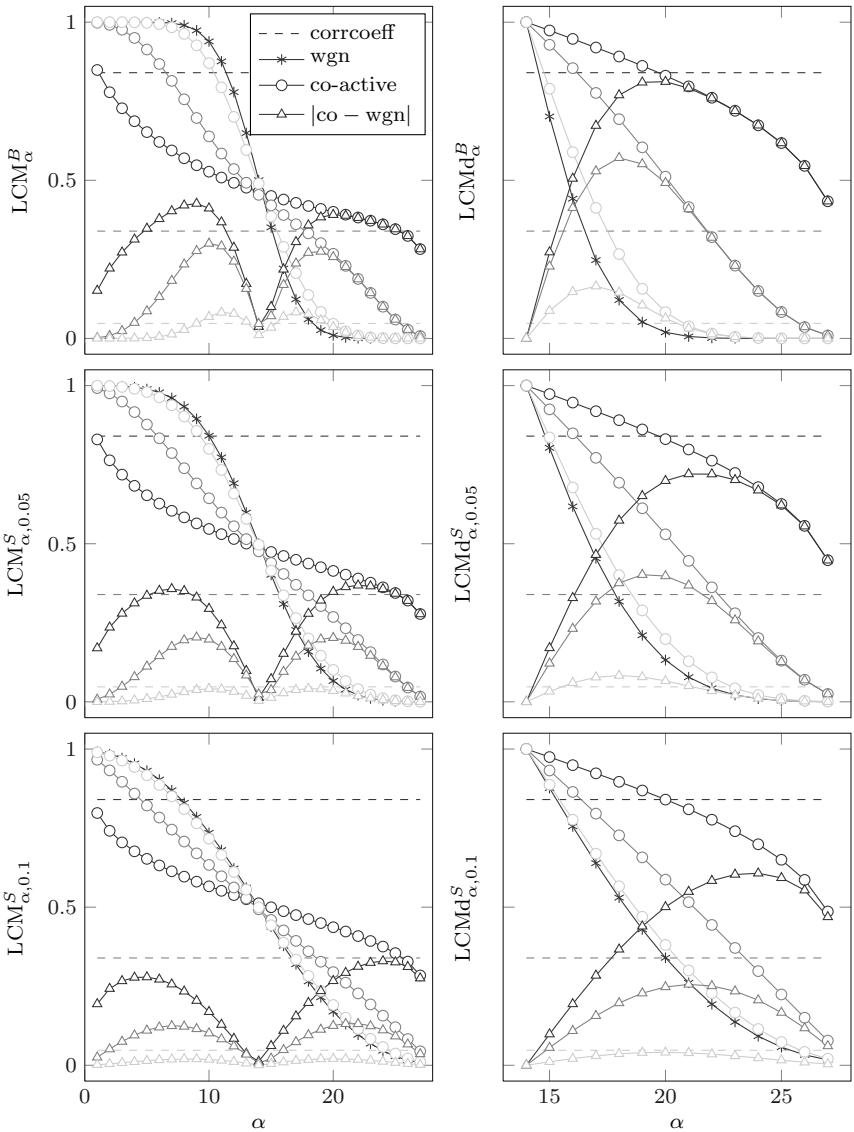


Fig. 2 The generated test data (Section 4) were subject to LCM^B_α (top left), LCMd^B_α (top right), $\mathrm{LCM}^S_{\alpha,0.05}$ (mid left), $\mathrm{LCMd}^S_{\alpha,0.05}$ (mid right), $\mathrm{LCM}^S_{\alpha,0.1}$ (bottom left), and $\mathrm{LCMd}^S_{\alpha,0.1}$ (bottom right). Sample means corresponding to the three data sets consisting of co-active 27-cuboids (circles) and to the WGN 27-cuboids (asterisks) were plotted against α. For the former, the darkness of the gray decreases with the SNR used when adding WGN to the time series. The absolute differences between the results corresponding to the WGN data and those corresponding to the three co-active data sets were plotted adopting the respective gray levels (triangles). The same holds for the horizontal dashed lines representing the mean of the average correlation coefficient (of the 351 pairwise correlations per sample) distribution of the respective co-active data set.

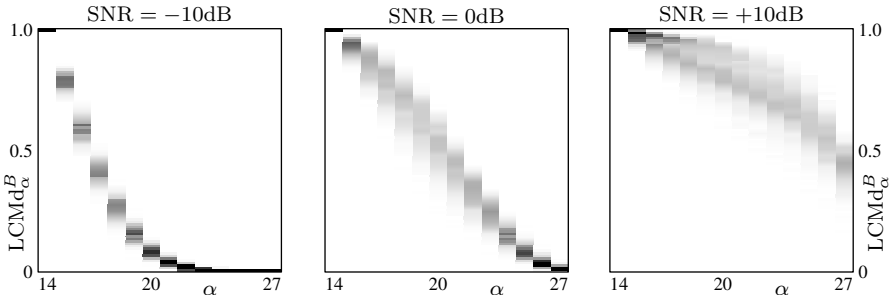


Fig. 3 Distribution of local connectivity measure LCMd for three signal-to-noise ratios.

The average correlation coefficient yields higher contrast between co-active and WGN samples (as estimated by the difference of their respective results) if the strength of the noise is low, that is, for higher SNRs. The opposite is true for lower SNRs: here LCM/LCMd outperforms the average correlation coefficient in the course of decreasing SNR. Unavoidably, however, the higher the strength of the additive WGN, the more the LCM/LCMd of the co-active samples resemble those of the WGN data.

As illustrated by Figure 3, the variance (and range) of LCMd—and therefore its sensitivity—is lowest for the most extreme values of α. Accordingly, also the difference between the co-active and the WGN voxels is minimal for $\alpha = 14$ and $\alpha = 27$ (Figure 2, right column). As explained in Section 3, the best choice of α depends on the level of noise present in the data and the desired contrast between functionally connected and functionally independent 27-cuboids. While a smaller α provides higher noise robustness (fault tolerance), a too small choice will impair the contrast between co-active voxels and WGN voxels, as the expected value under the assumption of independent voxels comes closers. Then again, some fault tolerance needs to be ensured, as due to noise a large α will result in a low range and similar LCMd for both co-active and noise voxels, all the same.

The attainable noise robustness of LCM seems to increase with the sharpening effect, that is, with decreasing slope β. The smaller β is, the farther each value in a time series gets shifted towards minimum or maximum, that is, towards 0 or 1. Thus, with β decreasing, LCM values tend to increase for $\alpha > 14$ and to decrease for $\alpha \leq 14$. In other words, the soft approach seems to keep more noise than actual information. However, this behavior may be due to our scheme of transforming the time series (using a median and quantile normalization) and further investigations are needed in order to clarify this.

6 Results for Real Data

The model-driven approach using a GLM yielded robust activation in cortical and subcortical visual regions mostly confined to gray matter. As expected, highest activity differences were found in the left and right calcarine sulcus respectively, the location of retinotopically organized primary visual cortex (V1). Additional regions representing the left and right visual field, respectively, could be localized such as the secondary and mid-level visual regions V2 adjacent to the dorsal and ventral part of V1 as well as area MT+, located bilaterally in the temporal parts of occipital cortex. Subcortical regions such as the lateral geniculate nucleus (LGN) as well as the superior colliculus showed statistically significant differences between their left and right visual field representation, albeit much lower than V1 (see Section 4).

We now set out to adress the question whether it is possible to generate meaningful brain maps without any *a priori* assumptions with respect to experimental design or hemodynamic properties across the brain. The LCM analysis generally yielded higher values for gray matter regions than for non-gray matter regions. As for SPM_t, the highest values were found in left and right V1 and adjacent visual regions. Visual inspection of the results indicated that high SPM_t values are most often accompanied by high LCM values (3^{rd} and 4^{th} row of Figure 4). The two-dimensional histogram of SPM_t and LCM confirmed this observation (5^{th} row of Figure 4). Conversely, many voxels were exhibiting no considerable activation associated with the visual experiment, while at the same time showing a high LCM. For both hemispheres, two regions of interest (ROI) of 100 voxels each were defined in the center of GLM activation, i.e., in V1, and in a white matter (WM) region in the temporal lobe, where no coherent—let alone visually driven—activity was to be expected. Highly active voxels in both V1 ROIs, identified with GLM and indicated by high SPM_t deviations from zero, were also identified by LCM, while both approaches identified the WM regions as non-responsive. Voxels identified as highly active by SPM_t also show high LCM values, suggesting local connectivity increases as the cortex is active (5^{th} row of Figure 4).

7 Discussion, Conclusions and Future Work

We presented noise-robust and data-driven measures that characterize local functional connectivity patterns in fMRI data. Specifically, the proposed LCMs are designed in order to capture the proportion of synchronous activity (LCM, $\alpha > 14$), synchronous inactivity ($1 - \mathrm{LCM}$, $\alpha < 14$) or both (LCMd) as exhibited by adjacent voxels during a fixed period of time.

Using fMRI data from a study of the visual path, we compared stimulus-related activity as detected by conventional regression-based GLM analysis with local functional connectivity as estimated by LCMd. While increased

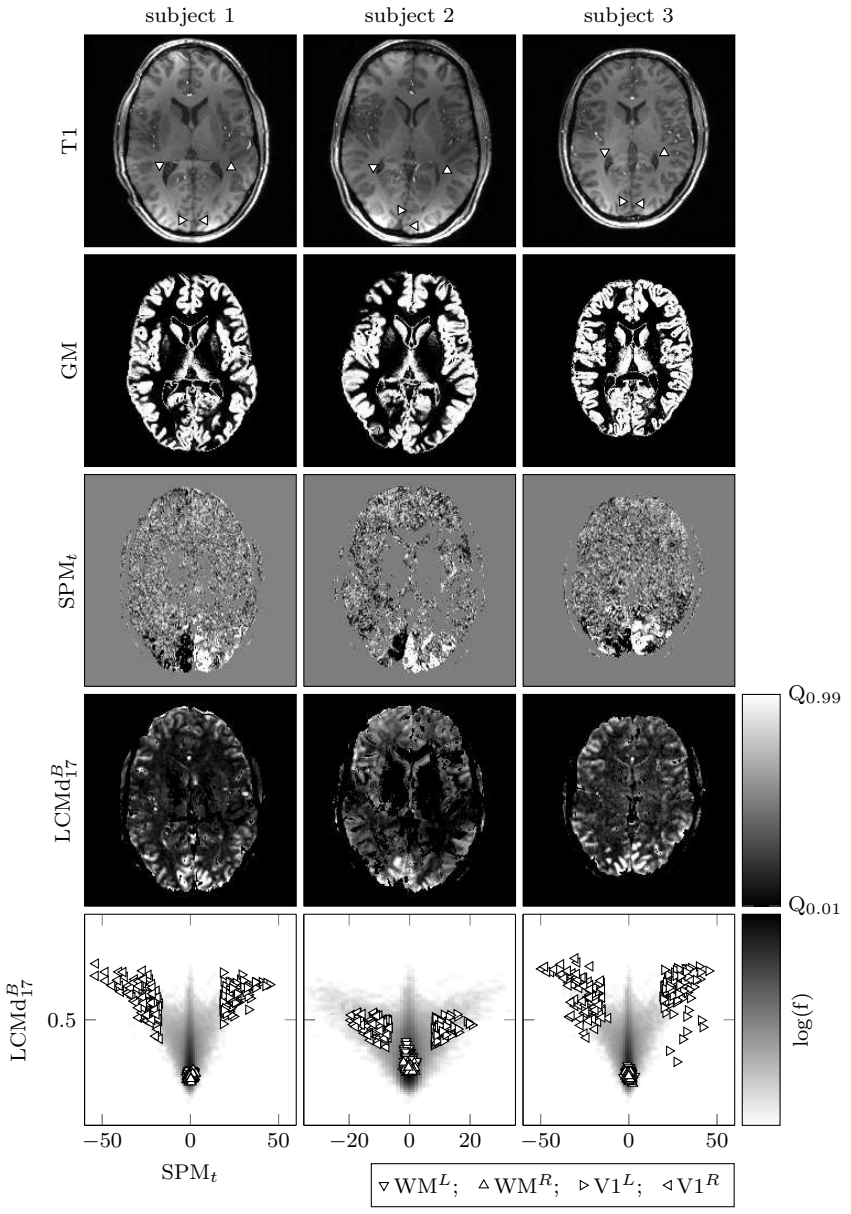


Fig. 4 GLM and LCM results. Each column corresponds to one subject from the visual stimulation experiment (Section 4). For each panel in row 1-4 the underlying data were clipped and scaled according to its respective 0.01- and 0.99-quantile before an axial slice was mapped to a gray color scale. 1st row: T1-weighted anatomical image. 2nd row: gray matter (GM) probability map. 3rd row: GLM results (SPM$_t$). Difference map between activity caused by ipsi- and contra-lateral visual field stimulation. 4th row: LCMd$^B_{17}$ map. 5th row: SPM$_t$ vs. LCMd$^B_{17}$. Background: 2D histogram of all brain voxels whose GM probability exceeded 0.5. Four ROIs of 100 voxels each are shown on top.

stimulus-related activity was most often accompanied by increased local functional connectivity, we also detected functionally connected clusters that exhibited no considerable task-related activity. These clusters may have been affected by locally coherent noise patterns or they may have been engaged in neuronal activity unrelated (or at least not linearly related) to the visual stimuli. In the latter case, further inspection of these clusters might give rise to new hypotheses, subsequently testable in a conventional fashion.

Beyond the initial proof of concept, the proposed approach may be utilized in future neuroscientific research as well as possible therapeutical implementations. As the method is entirely data-driven, it is applicable to *any* fMRI data (task- or stimulus-induced, resting-state, etc.). As such, LCM-based analyses may be especially suited to the analysis of resting-state fMRI data, as in this case no experimental task or stimulus onsets exist on which regressors for GLM and fitting of cHRFs could be based.

With a properly adjusted implementation, the LCM-versions based on dichotomous time series (that is, with a binary notion of activity) allow for time-efficient analysis of very large and very many data sets. This aspect might be exploitable for real-time fMRI (rtfMRI). The rtfMRI methodology aims at efficiently analyzing neuroimaging data in an online fashion (that is, concurrently with the data acquisition by the scanner), the results of which may govern the adaptation of experimental stimulation and the interaction with the subject. The feasibility of online analysis of complex emotional and cognitive states has recently been shown [7], while the future aim of such methods lies in therapeutic neurofeedback-based training after traumatic brain injury, cognitive stress or neurological pathology and will potentially culminate in brain machine interfaces [6]. In this application domain, changes in local functional connectivity could serve as an indicator of changing activity patterns, as suggested by the comparison of GLM/SPM$_t$ and LCM results.

In addition, LCM might serve as a filter in order to constrain the brain voxels to be analyzed further based on the functional images only or in addition to a T1-based gray matter segmentation. In favor of this idea it can be said that the LCM maps and the GM probability maps seemed to be highly conform (which is not surprising, though, since no neuronal activity is to be expected in non-GM areas). In fact, an initial and general reduction to informative parts of fMRI data before the actual analysis (whether assumption free or not) may constitute an interesting field of potential applications.

Future work includes finding a better way of mapping the intensity signal as picked up by the scanner to an activation degree, since the shortcomings of our current mapping do not allow us to fully exploit the advantages of a soft approach, which inherently is better suited to maintain all relevant information. Secondly, we are working on pertubation schemes to generate surrogate data that can be used to derive p-values for the detected local connectivity. Finally, we are in the process of extending our approach to a time-efficient analysis of spatially *unconstrained* connectivity, which is made possible by bit-vector representations of a binary notion of activity.

Acknowledgements The work presented in this paper was supported by Short-Term Scientific Mission (STSM) grant 9059 (Kristian Loewe) of COST Action IC0702.

References

1. Aguirre GK, Zarahn E, D'Esposito M (1998) The Variability of Human, BOLD Hemodynamic Responses. *Neuroimage* 8(4):360–369. Elsevier, Amsterdam, Netherlands
2. Cordes D, Haughton VM, Arfanakis K, Carew JD, Turski PA, Moritz CH, Quigley MA, Meyerand ME (2001) Frequencies Contributing to Functional Connectivity in the Cerebral Cortex in "Resting-state" Data. *American Journal of Neuroradiology* 22(7):1326–1333. American Society of Neuroradiology, Oak Brook, IL, USA
3. Eguiluz VM, Chialvo DR, Cecchi GA, Baliki M, Apkarian AV (2005) Scale-free Brain Functional Networks. *Physical Review Letters* 94:18102. American Physical Society, College Park, MD, USA
4. Friston KJ, Frith CD, Liddle PF, Frackowiak RS (1993) Functional Connectivity: The Principal-component Analysis of Large (PET) Data Sets. *Journal of Cerebral Blood Flow and Metabolism* 13(1):5–14. Nature Publishing Group, London, United Kingdom
5. Handwerker DA, Ollinger JM, D'Esposito M (2004) Variation of BOLD Hemodynamic Responses across Subjects and Brain Regions and Their Effects on Statistical Analyses. *Neuroimage* 21(4):1639–1651. Elsevier, Amsterdam, Netherlands
6. Hollmann M, Mönch T, Mulla-Osman S, Tempelmann C, Stadler J, Bernarding J (2008) A New Concept of a Unified Parameter Management, Experiment Control, and Data Analysis in fMRI: Application to Real-time fMRI at 3T and 7T. *Journal of Neuroscience Methods* 175(1):154–162. Elsevier, Amsterdam, Netherlands
7. Hollmann M, Rieger JW, Baecke S, Lützkendorf R, Müller C, Adolf D, Bernarding J (2011) Predicting Decisions in Human Social Interactions using Real-time fMRI and Pattern Classification. *PloS ONE*, 6(10):e25304. Public Library of Science, San Francisco, CA, USA
8. Logothetis NK, Wandell BA (2004) Interpreting the BOLD Signal. *Annual Review of Physiology* 66:735–769 Annual Reviews, Palo Alto, CA, USA
9. Ogawa S, Lee TM, Nayak AS, Glynn P (1990) Oxygenation-sensitive Contrast in Magnetic Resonance Image of a Rodent Brain at High Magnetic Fields. *Magnetic Resonance in Medicine* 14(1):68–78. J. Wiley & Sons, Chichester, United Kingdom
10. Smith AM, Lewis BK, Ruttimann UE, Ye FQ, Sinnwell TM, Yang Y, Duyn JH, Frank JA (1999) Investigation of Low Frequency Drift in fMRI signal. *Neuroimage* 9(5):526–533. Elsevier, Amsterdam, Netherlands
11. Tomasi D, Volkow ND (2010) Functional Connectivity Density Mapping. *Proceedings of the National Academy of Sciences of the USA* 107(21):9885. National Academy of Sciences, Washington, DC, USA
12. Tschukalin A (2011) Noninvasive Lokalisation von magno- und parvozellulären Anteilen des humanen CGL mittels Hochfeld-MRT. Bachelor Thesis. Dept. of Computer Science, Otto-von-Guericke Universität Magdeburg, Germany
13. Van den Heuvel MP, Stam CJ, Boersma M, Hulshoff Pol HE (2008) Small-world and Scale-free Organization of Voxel-based Resting-state Functional Connectivity in the Human Brain. *Neuroimage* 43(3):528–539. Elsevier, Amsterdam, Netherlands
14. Zaitsev M, Hennig J, Speck O (2004) Point Spread Function Mapping with Parallel Imaging Techniques and High Acceleration Factors: Fast, Robust, and Flexible Method for Echo-planar Imaging Distortion Correction. *Magnetic Resonance in Medicine* 52(5):1156–1166. J. Wiley & Sons, Chichester, United Kingdom

Fuzzy Clustering based on Coverings

Didier Dubois[1] and Daniel Sánchez[2,3]

Abstract In this paper we propose fuzzy coverings as a way to perform fuzzy clustering of data on the basis of a fuzzy proximity relation. Remarkably, the proposal does not require any kind of fuzzy transitivity.

1 Introduction

Clustering consists in finding a partition of a set of objects comprised of a collection of subsets of objects called *clusters*. It is expected that i) all elements in a cluster are *similar*, ii) every object belongs to a cluster, and iii) objects are in one cluster only or, equivalently, objects in different clusters are not similar.

The starting point for a clustering process is a similarity relation verifying the reflexivity, symmetry, and transitivity properties. If such a relation is available, clustering is just a matter of calculating the quotient set of the relation. However, several authors have pointed out that binary relations built in terms of natural concepts are not transitive most of the time, but reflexive and symmetric *indistinguishability relations*, also called *resemblance relations*. The situation is even worse in the case of fuzzy similarity relations [15, 12, 13], since any kind of fuzzy transitivity consists of transitivity at level 1 plus some additional requirements on the pairs of objects related to degrees in (0,1).

A natural consequence for the problem of clustering is that, in the crisp case, it is not always possible to obtain a partition of data according to the binary relation that represents the natural indistinguishability between objects. When indistinguishability is represented by a reflexive and symmetric,

[1] IRIT, CNRS & Université de Toulouse, France, dubois@irit.fr
[2] European Centre for Soft Computing, Mieres, Spain daniel.sanchezf@softcomputing.es
[3] Dept. Computer Science and AI, Universidad de Granada, Spain, daniel@decsai.ugr.es

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 319–330.
springerlink.com

but not transitive relation, what can be reasonably expected is to obtain not a partition, but a *covering* of the set of objects, in which an object may appear in more than one cluster. This is also the case in fuzzy clustering, though objects may appear in different clusters to a certain degree, and the result of the clustering process is called *fuzzy partition* and not *fuzzy covering*.

It is not the objective of this paper to discuss the notion of fuzzy partition, for which there is no unique definition, and how it might be different from the notion of *fuzzy covering* existing in the literature [5]. Our objective is to propose a new fuzzy clustering algorithm based on obtaining crisp coverings from crisp binary relations satisfying reflexivity and symmetry properties only.

The paper is organized as follows: we propose crisp coverings for clustering with crisp relations in Section 2. This procedure is extended to obtain fuzzy clusters as fuzzy coverings in Section 3, and we study the relationship to the work by Bezdek and Harris [2] in Section 4. Section 5 is devoted to briefly putting forward complexity issues of the proposal. Finally, Section 6 contains our conclusions and ideas for future work.

2 Coverings for Non-transitive Crisp Relations

When the indistinguishability relation between objects is not transitive, we can only expect to obtain a collection of clusters that form a covering of the data. However, this does not mean at all that any covering is a good clustering of the objects according to the relation R. We want a set of clusters $\{C_1, \ldots, C_m\}$ satisfying:

1. $C_i \times C_i \subseteq R \; \forall C_i$
2. The covering reflects *all* the information given by R in the following sense: let $O' \subseteq O$ such that $O' \times O' \subseteq R$. Then, there is at least one C_i such that $O' \subseteq C_i$, i.e., every group of objects that are completely related appear together in at least one cluster. This requirement also implies $\bigcup_{i=1}^{m}(C_i \times C_i) = R$.
3. There is no other clustering verifying 1 and 2 with less clusters.

It is easy to show the following proposition:

Proposition 1. *Let R be a reflexive and symmetric relation defined on a set O and let $G_R = (O, D)$ be a graph in which objects are the vertices and there is an edge in D between vertices x and y iff xRy. There is a single covering of O that yields a clustering satisfying criteria 1-3, comprised of the clusters corresponding to the sets of vertices of all the maximal cliques in G_R.*

Proof. Let C be a cluster, then $C \times C \subseteq R$ and there is a clique in G_R whose vertices are C. If the clique associated to C is a maximal clique of G_R then it is in the optimum covering. Otherwise, there is a maximal clique in G_R

with vertices C' so that $C \subset C'$ and C' is a cluster of the optimum covering. Hence, cluster C can be discarded since cluster C' is enough in order to satisfy criterion 2.

This proposition provides a procedure to obtain the best clustering by covering: we just need to calculate all the maximal cliques in the graph G_R. We shall consider this approach in the rest of this paper. This idea has been also proposed and employed before in the literature, like for example in [1, 9], and is also similar to finding concepts from Boolean matrices in Formal Concept Analysis [7]. The resulting covering is obviously biunivocally related to the relation R. The main difficulty with this approach is that the computation of all maximal cliques of a graph is an NP-complete problem. Hence, when the set of objects is very large, it may be the case that only approximate results can be achieved. We will come back to this problem in Section 5.

For instance, consider the non-transitive relation R represented in Table 1. The optimum clustering, formed by all the maximal cliques in the corresponding graph, is $\{\{a, b\}, \{b, c, d\}, \{c, d, e\}\}$.

	a	b	c	d	e
a	**1**	**1**	0	0	0
b	**1**	**1**	**1**	**1**	0
c	0	**1**	**1**	**1**	**1**
d	0	**1**	**1**	**1**	**1**
e	0	0	**1**	**1**	**1**

Table 1 Example of a crisp, reflexive, symmetric, and non-transitive relation R.

3 Coverings for Fuzzy Clustering

In this section we consider the problem of clustering objects on the basis of a fuzzy indistinguishability relation satisfying the reflexive and symmetric properties, i.e., a fuzzy relation R on O satisfying:

- $R(x, x) = 1 \ \forall x \in O$
- $R(x, y) = R(y, x) \ \forall x, y \in O$

3.1 Gradual Coverings

Our approach consists in solving the corresponding crisp clustering problem for each α-cut of R, obtaining a crisp covering at each level. This is possible since, as it is easy to show, α-cuts of fuzzy reflexive and transitive relations

are reflexive and symmetric crisp relations, so we are in the case discussed in the previous section.

Let Λ_R be the collection of significant levels in $[0,1]$ considered and let ρ_{C_R} be a function assigning to every value $\alpha \in \Lambda_R$ the optimum covering of O following the crisp relation R_α. We call the pair (Λ_R, ρ_{C_R}) a *gradual covering* (or *RL-covering*, following the notion of *representation by levels* of membership introduced in [10] and akin to the gradual entities proposed in [6]). The idea of representation by levels is that of assigning crisp representatives to membership levels, without imposing any restriction on the relationship between representatives on different levels, and then operating on each level independently.

For example, consider the fuzzy relation E in Table 2, proposed in [2]:

	a	b	c	d
a	**1**	0.3	0.6	0
b	0.3	**1**	0.7	0
c	0.6	0.7	**1**	0.3
d	0	0	0.3	**1**

Table 2 Fuzzy relation E introduced in [2].

Then $\Lambda_E = \{1, 0.7, 0.6, 0.3\}$. Table 3 shows the crisp relations corresponding to the α-cuts E_α with degrees $\alpha \in \Lambda_E$ and the corresponding optimum coverings on each level. The result is a gradual covering for E.

α	**Crisp relation** E_α	**Covering** $\rho_{C_E}(\alpha)$
1	$\begin{matrix}\mathbf{1}\,0\,0\,0\\0\,\mathbf{1}\,0\,0\\0\,0\,\mathbf{1}\,0\\0\,0\,0\,\mathbf{1}\end{matrix}$	$\{a\},\{b\},\{c\},\{d\}$
0.7	$\begin{matrix}\mathbf{1}\,0\,0\,0\\0\,\mathbf{1}\,\mathbf{1}\,0\\0\,\mathbf{1}\,\mathbf{1}\,0\\0\,0\,0\,\mathbf{1}\end{matrix}$	$\{a\},\{b,c\},\{d\}$
0.6	$\begin{matrix}\mathbf{1}\,0\,\mathbf{1}\,0\\0\,\mathbf{1}\,\mathbf{1}\,0\\\mathbf{1}\,\mathbf{1}\,\mathbf{1}\,0\\0\,0\,0\,\mathbf{1}\end{matrix}$	$\{a,c\},\{b,c\},\{d\}$
0.3	$\begin{matrix}\mathbf{1}\,\mathbf{1}\,\mathbf{1}\,0\\\mathbf{1}\,\mathbf{1}\,\mathbf{1}\,0\\\mathbf{1}\,\mathbf{1}\,\mathbf{1}\,\mathbf{1}\\0\,0\,\mathbf{1}\,\mathbf{1}\end{matrix}$	$\{a,b,c\},\{c,d\}$

Table 3 Gradual covering for fuzzy relation E of Table 2

Notice that, on each level α_i, the crisp covering $\rho_{C_R}(\alpha_i)$ and the crisp relation R_{α_i} corresponding to the α_i-cut of R are biunivocally related and

can be obtained one from the other, as we saw in the previous section. In addition, the fuzzy relation R and its level representation by α-cuts are also biunivocally related by the representation theorem of fuzzy sets. Hence, so are R and the gradual covering (Λ_R, ρ_{C_R}).

3.2 Fuzzy Clustering

Fuzzy clustering algorithms yield a so-called fuzzy partition as a result, consisting of a collection of fuzzy subsets of O verifying certain properties. There is no unique definition of the concept of fuzzy partition, but something common to all approaches is that the cores of the different fuzzy clusters are pairwise disjoint. This idea implies that the 1-cut of the fuzzy relation is a crisp relation verifying transitivity.

In our case, we are not assuming transitivity at all in our fuzzy relation, so it may be the case that the cores of the fuzzy clusters form a covering of O, but not a partition. Hence, fuzzy clustering based on a non-transitive (in any sense) fuzzy relation in general yields a *fuzzy covering*, i.e., a collection of fuzzy clusters $\{C_1, \ldots, C_m\}$ verifying $O = \bigcup_{i=1}^{m} C_i$ via some t-conorm [5]. Our objective in this section is to show that it is possible to derive a fuzzy covering from a gradual covering so that the former describes all the information in the latter, and they are biunivocally related. For that purpose, let us first introduce the notion of gradual cluster as follows:

Definition 1. Let (Λ_R, ρ_{C_R}) be a gradual covering for a fuzzy relation R. Let $\Lambda_R = \{\alpha_1, \ldots, \alpha_k\}$ with $1 = \alpha_1 > \alpha_2 > \cdots > \alpha_k > \alpha_{k+1} = 0$. Let $\rho_{C_R}(\alpha_i) = \{C_{i1}, \cdots, C_{in_i}\}$ be the set of clusters forming the optimum covering of R_{α_i}. A gradual cluster of (Λ_R, ρ_{C_R}) is a pair $Z = (\Lambda_Z, \rho_Z)$ defined by:

- $\Lambda_Z = \Lambda_R$
- $\rho_Z(\alpha_i) = C_{ip_i}$ with $1 \leq p_i \leq n_i$ satisfying $\rho_Z(\alpha_i) \subseteq \rho_Z(\alpha_{i+1})$ $\forall 1 \leq i \leq k-1$

That is, a gradual cluster can be obtained from a gradual covering by picking up one single crisp cluster from each level under the restriction that the clusters are nested according to the levels like α-cuts of fuzzy sets. Notice that, given a cluster in level α_i, it is always possible to find a cluster in level α_{i+1} such that the former is a subset of the latter since $R_{\alpha_i} \subset R_{\alpha_{i+1}}$ and hence a maximal clique of R_{α_i} is either a maximal clique of $R_{\alpha_{i+1}}$, or it is strictly included in a maximal clique of $R_{\alpha_{i+1}}$. By construction consecutive families $\rho_{C_R}(\alpha_i)$ and $\rho_{C_R}(\alpha_{i+1})$ satisfy a generalised containment property in the sense that

- $\forall C_{il} \in \rho_{C_R}(\alpha_i), \exists C_{i+1,j} \in \rho_{C_R}(\alpha_{i+1})$ such that $C_{il} \subseteq C_{i+1,j}$
- $\forall C_{i+1,j} \in \rho_{C_R}(\alpha_{i+1}), \exists C_{il} \in \rho_{C_R}(\alpha_i)$ such that $C_{il} \subseteq C_{i+1,j}$

So we can consider the graph made of all clusters at all levels, with arcs joining all $C_{il} \in \rho_{C_R}(\alpha_i)$ to all $C_{ij} \subseteq C_{i+1,j}$ whenever $C_{il} \subseteq C_{i+1,j}$. This graph is a directed acyclic graph which is unique since the set of clusters at each level is uniquely defined. Gradual clusters are then in one-to-one correspondence with paths of nested clusters from level 1 to level k in the graph. When the fuzzy relation is an equivalence relation this graph becomes a standard Hasse diagram.

As an example, the set of gradual clusters for the gradual covering in Table 3 is shown in Table 4.

α	**RL-Cluster 1**	**RL-Cluster 2**	**RL-Cluster 3**	**RL-Cluster 4**
1	$\{a\}$	$\{b\}$	$\{c\}$	$\{d\}$
0.7	$\{a\}$	$\{b, c\}$	$\{b, c\}$	$\{d\}$
0.6	$\{a, c\}$	$\{b, c\}$	$\{b, c\}$	$\{d\}$
0.3	$\{a, b, c\}$	$\{a, b, c\}$	$\{a, b, c\}$	$\{c, d\}$

Table 4 Gradual clusters (RL-clusters) for the gradual covering in Table 3

In this particular case, the number of clusters coincides with the number of objects, but this is not true in general, as we shall see later with another example.

Proposition 2. *There is a one-to-one relation between a gradual covering* (Λ_R, ρ_{C_R}) *and the set* $Z(R)$ *containing* all *the possible gradual clusters that can be obtained from* (Λ_R, ρ_{C_R}) *according to definition 7.*

Proof. Given a gradual covering (Λ_R, ρ_{C_R}) there is obviously a single set $Z(R)$ formed by all possible gradual clusters that can be obtained from (Λ_R, ρ_{C_R}) according to definition 1. On the other hand, for every pair (α_i, C_{ij}) with $\alpha_i \in \Lambda_R$ and $C_{ij} \in \rho_{C_R}(\alpha_i)$ a crisp cluster, there is at least one gradual cluster $Z \in Z(R)$ such that $C_{ij} \in \rho_Z(\alpha_i)$ since there is at least one cluster $C \in \rho_{C_R}(1)$ such that $C \subseteq C_{ij}$. Hence, $\bigcup_{Z \in Z(R)} \rho_Z(\alpha_i) = \rho_{C_R}(\alpha_i)$, and hence $\bigcup_{Z \in Z(R)}(\Lambda_R, \rho_Z) = (\Lambda_R, \rho_{C_R})$ with the union restricted to levels with the same α proposed in [10].

Now, obtaining a fuzzy covering $F(R)$ from the set of gradual clusters $Z(R)$ is straightforward: each gradual cluster $Z \in Z(R)$ corresponds to the representation by levels of a fuzzy subset of O, which is a fuzzy cluster that we shall denote $\mu_Z \in F(R)$. It is easy to show that:

- Since level 1 is a covering of O, every object in O appears with degree 1 in at least one fuzzy cluster.
- All fuzzy clusters $\mu_Z \in F(R)$ are normalized fuzzy subsets of O
- For every $\mu_Z \neq \mu'_Z \in F(R)$ it is neither $\mu_Z \subset \mu'_Z$ nor $\mu'_Z \subset \mu_Z$
- The set of fuzzy clusters forms a fuzzy covering of O that is our fuzzy clustering on the basis of R. Particularly, for any t-conorm

$$\bigcup_{F(R)} \mu_Z = O$$

- μ_Z is biunivocally related to Z, hence $F(R)$ is biunivocally related to $Z(R)$ and as a consequence, as we saw before, also to (Λ_R, ρ_{C_R}) and R.

In our example, the fuzzy clusters corresponding to the gradual clusters in Table 4 are the following:

1. $1/a + 0.3/b + 0.6/c$
2. $0.3/a + 1/b + 0.7/c$
3. $0.3/a + 0.7/b + 1/c$
4. $0.3/c + 1/d$

Notice also that, though rather similar in this example, these fuzzy clusters do not correspond exactly to the columns of the matrix defining the fuzzy relation E. In particular, the fuzzy cluster 3 differs from the third column/row in the mentioned matrix since the membership for a is 0.3, whilst in the third column of the matrix it is 0.6. In the general case, this difference can be even more significant since the number of fuzzy clusters can be larger than the number of columns, as we shall see later with another example.

4 Relation to the Study by Bezdek and Harris

In [2], Bezdek and Harris study fuzzy partitions and relations with special emphasis on the role of different notions of transitivity, and the possibility of finding crisp partitions from a fuzzy relation by means of convex decompositions. In this section we discuss these issues in relation to our proposal.

4.1 Transitivity

Transitivity is not uniquely defined for fuzzy relations. Existing definitions are based on considering a t-norm $\wedge$ and a t-conorm $\vee$ as follows: a fuzzy relation R is $\wedge - \vee$ transitive iff $\forall x, y \in O$, $R(x, y) \geq \vee_{z \in O}(R(x, z) \wedge R(z, y))$. It is usual to consider $\vee = \max$, and common types of fuzzy transitivity are max-min, max-prod, and max-Lukasiewicz (called max-$\triangle$ in [2]). As t-norms are ordered, the different types of transitivity induce an ordering of the classes of fuzzy relations satisfying them in terms of inclusion, e.g., a fuzzy relation satisfying max-min transitivity verifies any other kind of transitivity, etc.

About the role of fuzzy transitivity in fuzzy clustering, it is well known that max-min transitive relations allow us to obtain a crisp hierarchical clustering comprised of nested partitions, since every α-cut of the relation is a crisp equivalence relation. This does not hold for other types of transitivity. In addition, let R be a fuzzy relation which is max-$\wedge$ transitive with $\wedge$ any t-norm. Then, the kernel R_1 of R is a crisp equivalence relation. This is the

case for instance with relation E of Table 2, which verifies max-Lukasiewicz transitivity [2].

With respect to our approach to fuzzy clustering in relation to transitivity, we can say that:

- We do not require any kind of transitivity to hold for R.
- If any kind of transitivity holds, then it is guaranteed that i) the level 1 of the gradual covering is a crisp partition of O, and ii) the columns/rows for objects that are similar with degree 1 are equal, so they can be considered as a single element; in this case, we can eliminate redundant elements by a change in granularity of the problem.
- The relation R is max-min transitive iff our procedure (equivalent in this case to the usual procedure employed with this kind of relations) yields crisp partitions on each level, these partitions being nested in the usual way. The resulting gradual covering is then in fact a gradual partition (or RL-partition, see [10] for a definition).
- We conjecture: the more strict the transitivity that holds for R, the less the amount of overlap between clusters in the covering.

Let us now consider another example: relation S in Table 5, that does not satisfy any kind of transitivity.

	a	b	c	d	e
a	**1**	**1**	0	0.8	0.3
b	**1**	**1**	0.8	**1**	0.3
c	0	0.8	**1**	0.8	0.8
d	0.8	**1**	0.8	**1**	0.5
e	0.3	0.3	0.8	0.5	**1**

Table 5 Example of a reflexive, symmetric, and non-transitive fuzzy relation S.

Then $\Lambda_S = \{1, 0.8, 0.5, 0.3\}$. Table 6 shows the crisp relations corresponding to the α-cuts S_α with degrees $\alpha \in \Lambda_S$ and the corresponding optimum coverings on each level. The result is the gradual covering (Λ_S, ρ_{C_S}).

Table 7 shows all the gradual clusters for the gradual covering (Λ_S, ρ_{C_S}) in Table 6. Notice that there are six gradual clusters whilst there are only four crisp clusters in the covering at level 1, and only five objects, that is, the number of clusters is related neither to the number of objects nor the number of clusters in the covering at level 1. As indicated in Section 3.2, the gradual clusters correspond to every possible choice of one set from each level that respect the usual set inclusion of α-cuts with respect to α.

The fuzzy clusters corresponding to the gradual clusters in Table 7 are the following:

1. $1/a + 1/b + 0.8/d + 0.3/e$
2. $0.8/a + 1/b + 1/d + 0.3/e$
3. $1/b + 0.8/c + 1/d + 0.3/e$
4. $0.8/b + 1/c + 0.8/d + 0.3/e$
5. $0.3/b + 1/c + 0.5/d + 0.8/e$
6. $0.3/b + 0.8/c + 0.5/d + 1/e$

α	**Crisp relation** S_α	**Covering** $\rho_{C_S}(\alpha)$
1	$\begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix}$	$\{a, b\}, \{b, d\}, \{c\}, \{e\}$
0.8	$\begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \end{bmatrix}$	$\{a, b, d\}, \{b, c, d\}, \{c, e\}$
0.5	$\begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix}$	$\{a, b, d\}, \{b, c, d\}, \{c, d, e\}$
0.3	$\begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix}$	$\{a, b, d, e\}, \{b, c, d, e\}$

Table 6 Gradual covering for fuzzy relation S of Table 5

α	**RL-Clust. 1**	**RL-Clust. 2**	**RL-Clust. 3**	**RL-Clust. 4**	**RL-Clust. 5**	**RL-Clust. 6**
1	$\{a, b\}$	$\{b, d\}$	$\{b, d\}$	$\{c\}$	$\{c\}$	$\{e\}$
0.8	$\{a, b, d\}$	$\{a, b, d\}$	$\{b, c, d\}$	$\{b, c, d\}$	$\{c, e\}$	$\{c, e\}$
0.5	$\{a, b, d\}$	$\{a, b, d\}$	$\{b, c, d\}$	$\{b, c, d\}$	$\{c, d, e\}$	$\{c, d, e\}$
0.3	$\{a, b, d, e\}$	$\{a, b, d, e\}$	$\{b, c, d, e\}$	$\{b, c, d, e\}$	$\{b, c, d, e\}$	$\{b, c, d, e\}$

Table 7 Gradual clusters (RL-clusters) for the gradual covering in Table 6

4.2 Convex Decompositions

In [2], Bezdek and Harris study convex decompositions of fuzzy relations as a way to determine possible crisp partitions. They conclude that it is not always possible to find a convex decomposition, even when the fuzzy relation verifies certain kinds of transitivity like max-Lukasiewicz transitivity.

However, from our results in the previous section, it is immediate that it is always possible to find a convex decomposition of a fuzzy indistinguishability relation using crisp coverings, even when the fuzzy relation does not satisfy any kind of transitivity.

Proposition 3. *Consider a fuzzy relation R and let (Λ_R, ρ_{C_R}) be the optimum gradual covering of R following our approach. Let $\Lambda_R = \{\alpha_1, \ldots, \alpha_k\}$ with $1 = \alpha_1 > \alpha_2 > \cdots > \alpha_k > \alpha_{k+1} = 0$. Then*

$$R = \sum_{\alpha_i \in \Lambda_R} (\alpha_i - \alpha_{i+1}) \bigcup_{C \in \rho_{C_R}(\alpha_i)} C \times C \tag{1}$$

is a convex decomposition of R in terms of coverings.

The decomposition above is just a particular case of a well-known representation theorem for fuzzy sets with a finite level-set, and is employed in [10] as a way to obtain a fuzzy set from a representation by levels. This decomposition can be interpreted as the possible crisp coverings representative of the fuzzy relation, with associated importance degrees, that can be also seen as a basic probability assignment in the space of possible coverings and, hence, as a particular kind of random covering. Again, this decomposition is in a one-to-one relationship to R.

As examples, the convex decomposition for relation E, obtained on the basis of the gradual covering in Table 3 is (we show the relations, that are in one-to-one correspondence with their optimum coverings)

$$E = 0.3 \times \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} + 0.1 \times \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & \mathbf{1} & 0 \\ 0 & \mathbf{1} & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} + 0.3 \times \begin{bmatrix} \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & \mathbf{1} & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} + 0.3 \times \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & 0 & \mathbf{1} & \mathbf{1} \end{bmatrix}$$

and for relation S, the convex decomposition obtained on the basis of the gradual covering in Table 6 is

$$S = 0.2 \times \begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix} + 0.3 \times \begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \end{bmatrix} + 0.2 \times \begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix} + 0.3 \times \begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix}$$

5 Complexity

The main difficulty with the presented approach to fuzzy clustering based on coverings is that finding all the maximal cliques of a graph is an NP-complete problem, that cannot be even approximated in polynomial time [8]. Hence, this procedure is feasible if the size of the set O allows to compute the maximal cliques in acceptable time for the desired application.

There are a number of algorithms in the literature that try to compute maximal cliques as efficiently as possible [11, 14, 3]. The fact that for our purposes these algorithms have to be applied to every α-cut of the relation may be alleviated since each maximal clique at one level always includes maximal cliques in the previous level, and each clique at the latter level is included in at least one clique at the former level, hence allowing for bounding the subsequent search. Existing algorithms may be adapted in this sense. The time efficiency of the algorithms described in [11, 14, 3] can be found in the papers for a number of experiments. For instance, in [11], for a graph with 10^4 objects and different edge probabilities, finding all maximal cliques (up

to 230×10^6) took less than one hour in a Pentium 4 2.2Ghz computer. Further ideas about the kind of problems that may be approached using exact solutions can be obtained from the abovementioned papers.

As we mentioned before, the problem of clustering by coverings from crisp relations has been considered previously in the literature, and there are several proposals that try to find a covering by relaxing the requirement that clusters correspond to maximal cliques [1]. Applying these techniques to each level in order to obtain crisp coverings is a possibility in order to have an efficient algorithm, though properties like the one-to-one relationship between the fuzzy relation and the final clustering, and the idea of considering all clusters put forward by the fuzzy relation, are lost. A possibly better solution may be not to compute all maximal cliques, but a collection of maximal cliques such that they form a covering of the relation on each level. This way, we may lose some information in the form of some fuzzy clusters in the final solution, but we can obtain a subset of clusters that guarantee to cover the set of objects, from which the original relation can be recovered (though there is a one-to-many correspondence between the relation and such clusterings in general). We will explore the feasibility of the different possibilities discussed in this section in future works.

6 Conclusions and Future Work

When transitivity does not hold for a similarity relation, one cannot expect a clustering of objects in the form of a partition, but as a covering. This is also true in the case of fuzzy relations. We have proposed a way to obtain an optimum covering from a crisp reflexive and symmetric relation, which is unique and in a one-to-one correspondence with the original relation, based on the calculus of all maximal cliques of the relation. We have extended this procedure to the case of fuzzy relations, again without requiring any kind of fuzzy transitivity. Both representation by levels of a fuzzy covering, and a fuzzy clustering consisting in a fuzzy covering with fuzzy clusters, are proposed which are again in one-to-one correspondence to the original fuzzy relation. We have also studied the relation of the proposal to the study on fuzzy transitivity and convex decompositions by Bezdek and Harris [2] and we have shown that it is always possible to obtain a convex decomposition of a fuzzy relation in terms of crisp reflexive and symmetric relations and, therefore, in terms of the corresponding optimum coverings.

Calculating maximal cliques is NP-complete. Future work will be to study the situations in which current algorithms can be applied, as well as algorithms to obtain good approximations to the optimum solution in reasonable time when the problem is too large. We shall also study the relation to clustering with missing information, which has been studied previously by using coverings for representing ill-known partitions [4].

Acknowledgements This work has been supported by the COST Action IC0702 - Soft-Stat Combining Soft Computing Techniques and Statistical Methods to Improve Data Analysis Solutions - and by the Spanish Government under project TIN2009-08296.

References

1. Aslam J, Pelekhov K, Rus D (1998) Static and dynamic information organization with star clusters. *Proc. 7th Int. Conf. on Information and Knowledge Management*
2. Bezdek JC, Harris JD (1978) Fuzzy partitions and relations: An axiomatic basis for clustering. *Fuzzy Sets and Systems* 1(2):111–127
3. Cheng J, Ke Y, Fu AWC, Yu JX, and Zhu L (2010) Finding maximal cliques in massive networks by H*-graph. *Proc. 2010 Int. Conf. on Management of Data (SIGMOD'10)*
4. Couso I, Dubois D (2011) Rough sets, coverings and incomplete information. *Fundamenta Informaticae* 108:223–247
5. Dubois D, Prade H, eds. (2000) *Fundamentals of Fuzzy Sets.* Kluwer, Amsterdam
6. Dubois D, Prade H (2008) Gradual elements in a fuzzy set. *Soft Computing* 12:165–175
7. Dubois D, Prade H (2011) Bridging gaps between several frameworks for the idea of granulation. *Symposium on Foundations of Computational Intelligence (FOCI 2011)*, 59–65
8. Lund C, Yannakakis M (1994) On the hardness of approximating minimization problems. *Journal of the ACM* 41:960–981
9. Mohseni-Zadeh S, Brézellec P, Risler JL (2004) Cluster-c, an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. *Computational Biology and Chemistry* 28:211–218
10. Sánchez D, Delgado M, Vila MA, and Chamorro-Martínez J (2012) On a non-nested level-based representation of fuzziness. *Fuzzy Sets and Systems* 192:159–175
11. Tomita E, Tanaka A, and Takahashia H (2006) The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363:28–42
12. Trillas E, Valverde L (1984) An inquiry into indistinguishability operators. In: Skala HJ, Termini S, Trillas E, eds. *Aspects of Vagueness*, 231–256. D. Reidel, Dordrecht
13. Valverde L (1985) On the structure of f-indistinguishability operators. *Fuzzy Sets and Systems* 17:313–328
14. Wu B, Yang S, Zhao H, and Wang B (2009). A distributed algorithm to enumerate all maximal cliques in MAPREDUCE. *Proc. 4th Int. Conf. on Frontier of Computer Science and Technology (FCST'09)*
15. Zadeh LA (1971) Similarity relations and fuzzy orderings. *Information Sciences* 3(2):177–200

Decision and Regression Trees in the Context of Attributes with Different Granularity Levels

Kemal Ince[1] and Frank Klawonn[2,3]

Abstract Most data mining algorithms assume that their input data are described by a fixed number of attributes and each attribute has a pre-defined domain of values. However, the latter assumption is often not realistic in the case of categorical attributes. Such attributes are often available in different levels of granurality. The coarsest level might just have two possible values that are split into more values in refined levels of granularity. Before applying a data mining algorithm, it is usually assumed that the domain expert for the data must choose for each attribute the appropriate level of granularity or that in tedious trial and error procedure the appropriate granularity levels are adapted. The problem of choosing suitable granularity levels is related, but not identical to feature selection, since the more refined granularity levels increase the risk of overfitting. In this paper, we propose methods for decision and regression trees to handle the problem of different granularity levels during the construction of the corresponding tree.

1 Introduction

Data mining algorithms usually assume that the data to be analysed are encoded in a flat data table. Rows in the table correspond to data instances and columns to attributes. It is assumed that categorical attributes have a fixed range of possible values in such a table. This is, however, in many real world applications not the case. One often has a choice of picking the level

[1] Volkswagen AG, Komponenten-Werkzeugbau, Gifhornerstr. 180, D-38037 Braunschweig, Germany, kemal.ince@volkswagen.de
[2] Department of Computer Science, Ostfalia University of Applied Sciences, Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany, f.klawonn@ostfalia.de
[3] Bioinformatics and Statistics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany, frank.klawonn@helmholtz-hzi.de

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 331–342.
springerlink.com

of granularity for categorical attributes and only with the chosen level of granularity the range of possible values for a categorical attribute becomes fixed. An example for different levels of granularity is an attribute describing the type of a product that has been produced or bought. A very coarse level of granularity could be simply be the binary attribute *food* or *non-food*. On a refined level, *food* could be further divided into drinks and edibles. Drinks could be further distinguished into alcoholic and non-alcoholic drinks. Dealing with attributes that have different levels of granularity is a routine in data warehousing where OLAP (Online analytical processing) technologies are extensively used [3]. But in other fields, the problem of handling different levels of granularity of attributes [5] seems to have been neglected for quite a while, especially in the area of data mining.

In this paper, we describe an approach how different levels of granularity can be directly incorporated into the construction of regression trees based on the minimum description length principle (MDL). Section 2 explains the problem of different levels of granularity in the context of data mining and its relation to feature selection in more detail. Section 3 provides the necessary background on MDL, regression trees and granularity levels. Our regression tree algorithm incorporating the selection of granularity levels is described in Section 4. Section 5 briefly explains how granularity levels can be incorparated in the construction decision trees.

2 Brute Force vs. an Integrated Approach to Granularity Level Selection

As already mentioned in the introduction, choosing a suitable granularity level for an attribute is related to feature selection. Feature selection refers to choosing a subset of attributes which – in the ideal case – contains only the relevant and no redundant attributes. Feature selection are usually classified into two categories (see for instance [1]):

- Filter methods that carry out feature selection before the actual data mining algorithm is run. Filter methods try to remove irrelevant and redundant attributes without taking into account the specific data mining model to applied in a later step.
- Wrapper methods select the attribute in connection with the data mining algorithm. The result of the data mining algorithm with different subsets of attributes is evaluated and the subset that yields the best result is chosen. Often a greedy strategy like starting with only one attribute and adding more attributes step by step is applied to find a suitable subset of attributes.

Wrapper methods usually require a higher computational complexity, since the data mining algorithm must be carried out multiple times. Wrapper meth-

ods also have to cope with two other problems connected to overfitting. First of all, as long as there are not too many attributes, larger subsets of attributes tend to give a better performance than smaller ones, although this might only be an effect of overfitting. Secondly, the performance measure – for instance the performance of the derived data mining model on a test data set – should differ from the overall performance measure for the final model. Otherwise, overfitting with respect to the test data set might happen.

One could apply feature selection techniques to granularity level selection. This concept is illustrated in Figure 1. This would, however, lead to the same problems mentioned above as for filter and wrapper methods. Even worse, there is, of course, always a high correlation between different levels of granularity of the same attribute.

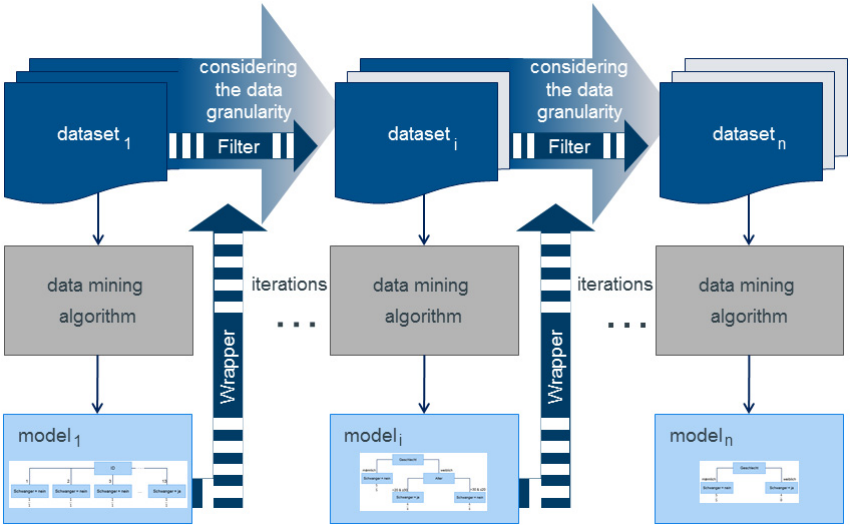


Fig. 1 Applying feature selection to granularity level selection.

Some data mining techniques carry out – at least partially – their own feature selection strategies while constructing the model. Decision and regression trees are examples for such methods. Their on-line selection of the attributes is based on the greedy strategy in which the tree is constructed. At each node of the tree during the construction, the attribute is chosen that leads to the best result with respect to a suitable performance measure.

In this paper, we focus on decision and regression trees to carry out granularity level selection during the construction of the tree. This integrated approach to granularity level selection is illustrated in Figure 2.

However, if we would simply allow the tree to choose in each of its nodes the attribute in connection with its level of granularity, there would be a strong bias to the most refined levels of granularity and therefore a tendency

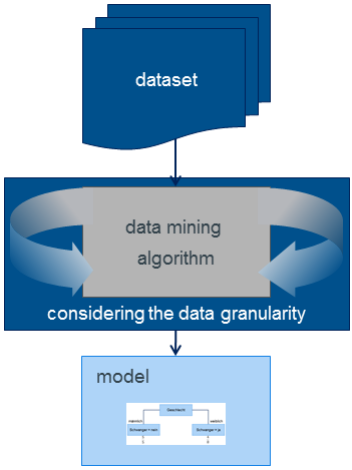


Fig. 2 New approach of the data mining analysis phase

to complex models and overfitting. Therefore, we need measures to avoid this bias.

3 Background on MDL, Regression Trees and Granularity Levels

3.1 The Minimum Description Length Principle (MDL)

Overfitting is a general problem in the context of fitting models[1] to data. Complex models tend to overfitting whereas too simple models will not be able to reflect the structure inherent in the data. What makes the situation complicated is that goodness of fit for models and model complexity are two different notions and they are measured in different "units". One way to try to minimise overfitting is to use different data for training the model and evaluating (or "testing") the models. A very common technique for model evaluation is cross-validation where models are trained and evaluated repeatedly with training and test data sets.

[1] We use the term *model* in a very broad sense as in [1]. Any structure that somehow describes some structure in the data is considered as a model: a single value like the mean, a linear model, a Gaussian mixture, a neural network or a decision tree are all considered as models.

The minimum description length principle (MDL) (for an overview see [4]) is an alternative approach to avoid overfitting. MDL is based on the fundemental idea that any regularity in a data set can be used to compress the data. Compression means to describe the data set with fewer symbols than the number of symbols which are needed to describe the data set literally. The description of the regularity is given by the model. The more regularities in the data set exist, the more the data set can be compressed. In this sense, MDL relates "learning" to "finding regularities" in the data.

Therefore, the MDL principle can be used in different ways of inductive inference such as to choose a model with a good trades-off between goodness-of-fit on the observed data set and the complexity of the model.

In this paper, we focus on supervised learning where the value of a target attribute is to be predicted on the basis of other predictor attributes. The compression of the data consists in this case of two parts.

- The coding of the model (here: the regression or decision tree) and
- the corrections for deviations of the value predicted by the model and the target value.

A very complex model – a complex regression or decision tree – will need a longer coding for the model itself, but no or little corrections for the predictions whereas a simple model has a short coding but requires a larger number or larger corrections for wrong predictions. MDL favours the model with the shortest coding of the model together with the corrections, so that a compromise between an extremely simple model with little precision and an overfitted complex model will be chosen.

3.2 Regression and Decision Trees

Regression and decision trees [2] are tree-based models for the prediction of numerical and categorical attributes. Each node in the tree corresponds to an attribute. Depending on the value of the attribute, the path is followed along the corresponding successor node which either corresponds to another attribute or to leave node which contains the predicted value.

Regression and decision trees are usually constructed by a greedy strategy. The root of the tree is the attribute that gives the improvement in terms of prediction. For decision trees, very often remaining entropy after splitting the data with respect to the attribute in the node is used as a measure for the predictive power of the attribute. The attribute with the smallest remaining entropy should be chosen as a root node. Successor nodes are treated in the same way until a unique prediction can be made, no attributes are left, not enough data end up in the corresponding node or some stop criterion like the maximum depth of the tree is reached. Regression trees are built in the same way, except that not the entropy, but the sum of the absolute or squared

errors in the successor nodes when the mean value is used as the predicted value is taken as a measure to judge the predictive quality of the attribute.

3.3 The Meaning of Data Granularity

As mentioned in the introduction, nowadays data sets to be analysed often contain attributes with different levels of granularity. Especially the multi-dimensional way of storing data, for example in data warehouses or online analytical processing systems, is one of the main reasons why handling data with different granularities is of high importance. The meaning of data granularities is illustrated in Table 1, where a fictitious data set contains the attributes A_1 and A_2 and consists of two granularity levels. For example the value a_1 of the attribute A_1 can have the two different specifications a_{11} and a_{12} in the refined attribute A_2. Attribute A_1 has a coarser level of granularity than attribute A_2. The target attribute Z contains of continuous numbers. This simple data set will serve as an illustrative example in the following section.

Table 1 Fictitious dataset containing two data granularity levels

A_1	A_2	Z
a_1	a_{11}	2
a_1	a_{11}	3
a_1	a_{12}	6
a_2	a_{21}	99
a_2	a_{21}	101
a_2	a_{21}	100
a_2	a_{22}	123
a_1	a_{12}	7
a_2	a_{22}	124
a_1	a_{11}	1
a_1	a_{12}	5
a_2	a_{22}	125

4 Constructing Regression Trees in the Context of Attributes with Different Levels of Granularity

Essentially, we can generate three different regression trees based on the simple data set in Table 1. Case one without any split delivers the regression tree shown in Figure 3. In this case there exists no splitting step, so that the root

node contains all data objects in the dataset and predicts the mean value of all data for the target attribute Z.

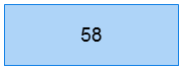


Fig. 3 Generated regression tree without splitting.

Case 1:
The prediction of the single node in terms of the mean value m_c in Figure 3 and the sum of squared errors S of the generated tree are calculated as follows:

$$m_c = \frac{1}{12} * (2+3+6+99+101+100+123+7+124+1+5+125) = 58 \quad (1)$$

$$\begin{array}{r} (2-58)^2 = 3136 \\ +(3-58)^2 = 3025 \\ +(6-58)^2 = 2074 \\ +(99-58)^2 = 1681 \\ +(101-58)^2 = 1849 \\ +(100-58)^2 = 1764 \\ +(123-58)^2 = 4225 \\ +(7-58)^2 = 2601 \\ +(124-58)^2 = 4356 \\ +(1-58)^2 = 3249 \\ +(5-58)^2 = 2809 \\ +(125-58)^2 = 4489 \\ \hline -->S = 35258 \end{array} \quad (2)$$

Case 2:
In the second case, the regression tree building step uses the predictor attribute A_1 to split the output values into two groups. Figure 4 shows the generated regression tree model. Therefore, the calculation of the prediction of both leaves m_{c1}, m_{c2} and the sum of squared errors with $S = S_1 + S_2 = 1004$ is given by

$$m_{c1} = \frac{1}{6} * (2+3+6+7+1+5) = 4 \quad (3)$$

$$m_{c2} = \frac{1}{6} * (99+101+100+123+124+125) = 112 \quad (4)$$

$$\begin{array}{rl|rl} (2-4)^2 & = 4 & (99-112)^2 & = 169 \\ +(3-4)^2 & = 1 & +(101-112)^2 & = 144 \\ +(6-4)^2 & = 4 & +(100-112)^2 & = 121 \\ +(7-4)^2 & = 9 & +(123-112)^2 & = 121 \\ +(1-4)^2 & = 9 & +(124-112)^2 & = 196 \\ +(5-4)^2 & = 1 & +(125-112)^2 & = 225 \\ \hline -->S_1 = 28 & & -->S_2 = 976 & \end{array} \quad (5)$$

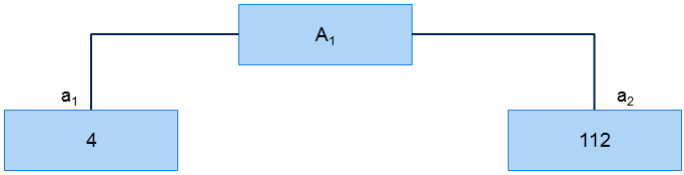


Fig. 4 Generated regression tree using A_1 for splitting.

Case 3:
The third case describes the selection of the refined predictor attribute A_2 for splitting delivers the regression tree with four leave nodes with predicted values m_{c11},m_{c12},m_{c21} and m_{c22} as visualied in Figure 5. The calculations are as folows.

$$m_{c11} = \frac{1}{3} * (2+3+1) = 2 \quad (6)$$

$$m_{c12} = \frac{1}{3} * (6+7+5) = 6 \quad (7)$$

$$m_{c21} = \frac{1}{3} * (99+101+100) = 100 \quad (8)$$

$$m_{c22} = \frac{1}{3} * (123+124+125) = 124 \quad (9)$$

The sum of squared errors $S = S_{11} + S_{12} + S_{21} + S_{22} = 8$ is calculated as follows.

$$\begin{array}{|-|(2-2)^2 = 0|-|(6-6)^2 = 0|-|(99-100)^2 = 1|-|+(123-124)^2 = 1 \\ |-|+(3-2)^2 = 1|-|+(7-6)^2 = 1|-|+(101-100)^2 = 1|-|+(124-124)^2 = 0 \\ |-|+(1-2)^2 = 1|-|+(5-6)^2 = 1|-|+(100-100)^2 = 0|-|+(125-124)^2 = 1 \\ |-|-->S_{11} = 2 \quad |-|-->S_{12} = 2 \quad |-|-->S_{21} = 2 \quad |-|-->S_{22} = 2 \end{array} \quad (10)$$

We can imagine that apart from the two predictor attributes A_1 and A_2, there are more predictor attributes. When we construct the regression tree, we have decide which attribute and in the case attributes A_1 and A_2 which level of granularity we choose for splitting in a node. If we only use the sum of squared errors, then the refined attribute A_2 will always be prefered over

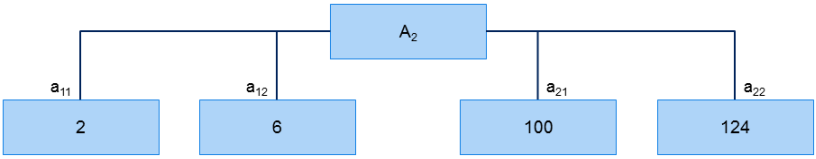


Fig. 5 Generated regression tree using A_2 for splitting.

the coarser attribute A_1, since the latter one can never yield a smaller sum of squared errors than A_2.

To avoid this effect, we apply the MDL principle during the splitting step to decide whether a split refined makes sense or not. The MDL measure MDL_{RT} of the generated regression tree is a binary value and consists of the sum of the number of digits N_{RT} which have to be corrected and the binary coding of the generated (partial) tree C_{RT} at the corresponding splitting node. In the following, both values are calculated for the three different cases. The number of digits which have to be corrected is calculated as

$$N_{RT} = \sum_{k=1}^{8} f_i \tag{11}$$

with $k = 1, ..., 8$ the positions of the binary digits of an 8 bit value[2] and $f = 1, ..., 12$ as the sum of digits to be corrected for data object i in the data set visualised in Table 1. The binary coding of the tree depends on the number of successor nodes. The splitting must be encoded and for each successor node a predicted value is needed. In Table 2 the values of the target attribute Z of Table 1 are shown as binary numbers. The following example illustrates how the calculations for MDL are carried out. The value of the third data object d_3 is $(6)_{10}$ with the binary coding $(00000110)_2$ containing 8 digits. In the first case, the prediction of the regression tree is $(58)_{10}$. In binary coding this is $(00111010)_2$. This means the number of digits which have to be corrected is $f_3 = 8$:

$$\begin{array}{rl} k = & 87654321 \\ & 00111010 = (6)_{10} \\ - & 00000110 = (58)_{10} \\ & \overline{11001100} = \overline{(52)_{10}} \end{array} \tag{12}$$

Case 1:

In this case the regression tree consists of one node as shown in Figure 3. Therefore, the prediction for every single data object in the data set is $(58)_{10}$ and in binary code 00111010. The binary calculation of N_{RT_1} delivers the

[2] Here we choose a precision of 8 bits. Of course, depending on the application, a higher or lower precision can be required.

Table 2 Output parameter values in binary code

$(z)_{10}$	$(z)_2$
2	00000010
3	00000011
6	00000110
99	01100011
101	01100101
100	01100100
123	01111011
7	00000111
124	01111100
1	00000001
5	00000101
125	01111101

following result:

$$N_{RT_1} = (8+8+8+6+6+6+7+8+7+8+8+7) = 87 \tag{13}$$

The generated tree RT_1 contains one single node with the binary coding $C_{RT_1} = 00000001$. The resulting MDL measure is $MDL_{RT_1} = N_{RT_1} + C_{RT_1}$ with $MDL_{RT_1} = 88$.

Case 2:

The second case refers to the tree visualised in Figure 4 including 3 nodes. Resulting from this tree, the binary coding is $C_{RT_2} = 00000011$. N_{RT_2} must be calculated for two different leaves of the generated tree and delivers the following result:

$$N_{RT_2} = N_{RT_{2_a}} + N_{RT_{2_b}} = 58 \tag{14}$$

$N_{RT_{2_a}}$ represents the case predicting 4 and $N_{RT_{2_b}}$ the case predicting 112. Both values were calculated as shown in the following.

$$N_{RT_{2_a}} = (2+2+3+3+2+1) = 13 \tag{15}$$

$$N_{RT_{2_b}} = (8+8+8+7+7+7) = 45 \tag{16}$$

The $MDL_{RT_2} = 58 + 3 = 61$ value is smaller for this tree.

Case 3:

The last and third case in the example is the regression tree shown in Figure 5 including 5 nodes. The binary coding resulting from these nodes is $C_{RT_3} = (00000101)_2$. The sum of the number of digits which have to be corrected N_{RT_3} in this case must be calculated by considering the different prediction values 2, 6, 100 and 124 of the tree.

$$N_{RT_3} = N_{RT_{3_a}} + N_{RT_{3_b}} + N_{RT_{3_c}} + N_{RT_{3_d}} = 8 \tag{17}$$

$$N_{RT_{3_a}} = (0 + 1 + 1) = 2 \tag{18}$$
$$N_{RT_{3_b}} = (0 + 1 + 1) = 2 \tag{19}$$
$$N_{RT_{3_c}} = (1 + 1 + 0) = 2 \tag{20}$$
$$N_{RT_{3_d}} = (1 + 0 + 1) = 2 \tag{21}$$

The last tree visualised in Figure 5 yields an even smaller MDL-value of $MDL_{RT_3} = 8 + 5 = 13$. Considering the MDL value as measurement implies that the regression tree in the third case delivers the best model for the original data set.

5 Decision Trees

For decision trees, there is another option than MDL to incorporate the selection of the most suitable attribute and its level of granularity at the same time. If we only focus on the information gain, i.e. the reduction of the entropy that is achieved by splitting with respect to a certain attribute, we would also prefer the most refined level of an attribute, since the entropy will always decrease with further splitting. However, for the construction of ordinary decision trees, there is already a technique that we can exploit. The gain ratio [6] was originally introduced to avoid the effect that predictor attributes having more different values tend to be prefered over attributes with few values by the simple information gain as a measure for deciding which attribute to use for a split in a node. Although the gain ratio was originally introduced to compare different attributes with a different number of values, it can also be applied directly to a single attribute with different levels of granularity.

6 Conclusions

In this paper we have addressed the problem of attributes with different levels of granularity in the context of data mining which is often neglected although it is a common challenge in real world applications. This problem is often tackled in a way as illustrated in Figure 1, leading to high computational costs when all combinations of granularity levels of different attributes are considered. At least for decision and regression trees, the computational costs can be reduced drastically by incorporating the selection of the granularity levels directly in the construction of the model as it is done by the methods described in this paper.

Further research work is needed to extend these ideas to other data mining methods than decision and regression trees.

References

1. Berthold MR, Borgelt C, Höppner F, Klawonn F (2010) *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*. Springer, London
2. Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and Regression Trees*. Wadsworth, Belmont CA
3. Chaudhuri S, Dayal U (1997) An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec.* 28:65–74
4. Grünwald PD (2007) *The Minimum Description Length Principle*. MIT Press, Cambridge
5. Pedrycz W, Skowron A, Kreinovich V, eds. (2008) *Handbook of Granular Computing*. Wiley, Chichester
6. Quinlan, JR (1986) *Induction of Decision Trees*. Springer, New York

Stochastic Convergence Analysis of Metaheuristic Optimisation Techniques

Nikos S. Thomaidis[1] and Vassilios Vassiliadis[1]

Abstract Commonly used metaheuristic optimisation techniques imbed stochastic elements into the selection of the initial population or/and into the solution-search strategy. Introducing randomness is often a means of escaping from local optima when searching for the global solution. However, depending on the ruggedness of the optimisation landscape and the complexity of the problem at hand, this practice leads to a dispersion of the reported solutions. Instead of relying on the best solution found in a set of runs, as is typical in many optimisation exercises, it is essential to get an indication of the expected dispersion of results by estimating the probability of converging to a "good" solution after a certain number of generations. We apply a range of statistical techniques for estimating the success probability and the convergence rate of popular evolutionary optimisation heuristics in the context of portfolio management. We show how this information can be utilised by a researcher to obtain a deeper understanding of algorithmic behaviour and to evaluate the relative performance of competitive optimisation schemes.

1 Introduction

Metaheuristic optimisation techniques, such as genetic algorithms, particle swarms and ant colonies, are constantly gaining attention in a variety of application fields. Nowadays, they are considered by many researchers as a promising alternative to traditional gradient-search methods that is better suited to the complexities of real-life applications (high-dimensional data, combinatorial explosion, non-differentiable functions, multiple local optima) [4, 12, 17].

[1] Management and Decision Engineering Laboratory, Department of Financial and Management Engineering, University of the Aegean, 41 Kountouriotou Str., 82100 Chios, Greece, {nthomaid,v.vassiliadis}@fme.aegean.gr

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 343–357.
springerlink.com

In all of the aforementioned metaheuristic paradigms, the exploration of the solution space is guided by random elements. Introducing randomness is a strategy to avoid premature convergence and local optima. Hence, even though in some of the runs the algorithm might get stuck in suboptimal regions, increasing the number of repetitions will boost a better exploration of the solution space and, eventually, lead to the global optimum. A downside of this practice is that it unavoidably leads to a higher dispersion of the reported results. Even when the algorithm is initialised from exactly the same population, it will most likely follow a different convergence path and the reported optimal solution may change from run to run. The degree of divergence in algorithmic behaviour depends, of course, on the morphology of the optimisation landscape and the complexity of the problem at hand. Still, the bottom-line is that we need a different approach to the evaluation and execution of metaheuristic schemes. Since we are dealing with randomness, statistical tools can play an important role to this end.

Statistical techniques are becoming increasingly popular in the area of metaheuristic optimisation. In the early years, researchers would routinely judge the relative superiority of each algorithm by comparing average scores over a set of runs on a single data set. However, nowadays, many recommend the use of statistical tests for pairwise or multi-wise comparisons of algorithmic behaviour (see e.g. [5] for an overview of parametric/nonparametric techniques). Still, on the application side, relatively little has been done in the direction of using statistical techniques for fine-tuning algorithmic parameters or determining the exploration/exploitation tradeoff. Gilli and Winker [6] set forth a range of probabilistic tools for analysing the convergence properties of heuristic techniques and the empirical distribution of algorithmic outcomes in each generation. They additionally show how this stochastic analysis can set the ground for determining an optimal allocation of CPU power between exploitation and exploration. Barrero et al. [1] investigate the convergence rate of genetic programming through the concept of *generation-to-success* (i.e. the number of generations guaranteeing convergence to the optimal region). They propose various techniques for estimating the statistical distribution of this random variable on a given optimisation task. In [13], we investigate the success rates of three stochastic optimisation techniques (simulated annealing, genetic algorithms, particle swarms) in the task of designing portfolios with certain enhancements over a benchmark financial index. [14] is a continuation of the study mentioned above, providing additional experimental evidence on the synthesis of optimal asset allocations and the speed at which each optimisation scheme converges to the optimum region.

The objective of this chapter is to picture the performance of common optimisation metaheuristics by means of an expanded set of statistical techniques. Along the guidelines of [6], we attempt to provide a "cloud" of outcomes that portray different aspects of the uncertainty associated with algorithmic behaviour in a portfolio optimisation task. Instead of following the common practice of reporting the average solution or the best solution found in a set

of runs, we analyse the dispersion of results by estimating the so-called *cumulative success probability*. In our context, this is defined as the probability of converging, in a given number of iterations, to a solution *within the range of the global optimum*. Given that the success rate is estimated from a relatively small number of independent runs, there is additional uncertainty as to the reliability of this estimate beyond the experimental data set. To address this issue, we propose a resampling technique for computing confidence intervals on the cumulative success probability, which does not rest on a parametric distribution (e.g. normal). We also derive a related measure of algorithmic performance that provides an estimation of the *worst expected outcome* that is likely to be observed in a single run. These concepts are applied to the evaluation of two popular metaheuristics, genetic algorithm and artificial ant colonies, in a portfolio optimisation context. As a case study, we consider the problem of selecting an optimal capital allocation among a set of stocks by placing an upper limit on the size of the portfolio (cardinality constraint).

The rest of the chapter is structured as follows: Sect. 2 discusses the formulation of the optimisation problem examined in this study. Sect. 3 presents the general architecture of the metaheuristic solvers, while Sect. 4 gives information on the sample data and the experimental setting. Different aspects of the statistical performance of metaheuristic techniques are analysed in Sect. 5–7. Sect. 8 concludes the chapter and discusses directions for further research.

2 Problem Formulation

The objective of our portfolio optimisation problem is to maximise a commonly used performance criterion, the *Sortino ratio* (SoR) [11]. This measures the *excess expected return* delivered by the portfolio (i.e. the expected return minus the return on the risk-free asset) per unit of *downside standard deviation* of returns. For a random variable X with probability density function $p_X(x)$, the downside standard deviation is $V_\tau(X) \equiv \sqrt{\int_{-\infty}^{\tau}(\tau - x)^2 p_X(x)dx}$, where τ is a negative threshold defined by the fund manager. As seen, the downside standard deviation is also a measure of dispersion but it only takes into account returns that fall below τ. In our case, the threshold is set equal to zero. So, the main interest of the portfolio manager is to maximise the expected return and also minimise the probability of observing large losses.

Apart from the objective function, the portfolio selection problem is equipped with several restrictions which are typical in this type of application:

- *Full investment* constraint, i.e. the available capital is fully invested in risky assets.

- *Floor and ceiling* constraint. A portfolio has to be well-balanced i.e. no single asset (or group of assets) must absorb a large proportion of the initial installment. This requirement is fulfilled by imposing a lower and upper limit on each asset weight in the final capital allocation.

Most fund managers select their holdings from a universe of assets belonging to a popular market benchmark, such as the S&P 500 or the Russell 3000 stock indices. This is because their skill is often benchmarked against the index. However, index constituents change from time to time and some of the assets taken into account in the calculation of the index are held in very small quantities. A fund manager may stray from investing in all member stocks and instead focus on holding small and manageable bundles of assets, with which he/she might be able to attain the investment goal. This practical necessity translates into an upper limit on the number of assets included in the portfolio, the so-called *cardinality constraint*.

A formulation of the portfolio optimisation problem encapsulating all the requirements set above is the following:

$$\underset{\boldsymbol{\omega}\in\mathbb{R}^N\ \mathbf{b}\in\{0,1\}^N}{\text{maximise}}\ SoR(\boldsymbol{\omega},\mathbf{b}) \equiv \frac{E[R_P(\boldsymbol{\omega},\mathbf{b})]-R_f}{V_0[R_P(\boldsymbol{\omega},\mathbf{b})]} \tag{1a}$$

subject to

$$\sum_{i=1}^{N}\omega_i = 1 \tag{1b}$$

$$\omega^f \leq \omega_i \leq \omega^c,\ i=1,...,N \tag{1c}$$

$$\sum_{i=1}^{N} s_i \leq K \leq N \tag{1d}$$

where $\boldsymbol{\omega} \equiv (\omega_1,\omega_2,...,\omega_N)'$ is the vector of portfolio weights, showing the fraction of capital invested in each asset; $\mathbf{s} = (s_1,s_2,...,s_N)'$ is a vector of indicator variables, taking the value 1 if money is put on the ith asset and 0 otherwise; N is the size of the investment universe; $E[R_P]$ is the expected portfolio return; R_f is the return on a risk-free bank account; $V_0[R_P]$ is the downside standard deviation; ω^f, ω^c is the lower and upper limit on asset weights and K is the portfolio cardinality. Note from (1b) that all w_i's sum up to one, reflecting the requirement that the available capital be fully invested, and the total number of assets with non-zero weight is less than K (constraint (1d)).

The portfolio-selection formulation discussed above is a typical case of a *mixed-integer nonlinear programming* problem, which is hard to solve even for small values of N. In fact, the search for optimal solutions expands along two dimensions:

a) finding a suitable basket of K assets.
b) deciding the optimal capital allocation among these assets.

One could possibly simplify the solution-search strategy by making a complete enumeration of all possible asset combinations of maximum size K and deploy, for each combination, a gradient-search technique to detect optimal portfolio weights. However, the computational time required for this solution strategy increases exponentially with the size of the problem (i.e. the problem is NP-complete). Fig. 1 plots the number of possible ways with which

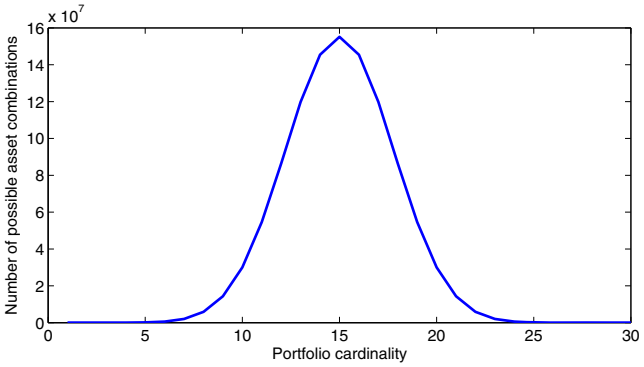


Fig. 1 Combinatorial complexity vs cardinality.

30 assets could be assigned into classes of size $K = 1, ..., 30$, where K is read on the horizontal axis. Note that even in such a relatively small investment universe, the number of feasible portfolio allocations literally explodes as we move towards medium-range cardinalities (i.e. when $K \approx N/2$).

3 Description of Optimisation Metaheuristics

Two stochastic metaheuristics are applied to the solution of the proposed portfolio-selection problem: a *genetic algorithm* (GA) and an *ant colony optimisation* (ACO) technique. We designed a *hybrid* solution-search strategy, whereby the aforementioned metaheuristics are only directed towards detecting promising combinations of K assets. For each step in the space of possible combinations, a Levenberg-Marquardt algorithm (see e.g. [10]) was run to determine optimal weights satisfying constraints (1b) and (1c). In what follows, we provide a summary presentation of metaheuristic techniques; more implementation details are given in the references below, in Sect. 4 and are also available form the authors upon request.

The genetic algorithm was firstly proposed by Holland [7] and draws on concepts from the process of biological evolution. The idea is to reach a pop-

ulation of "high-quality" solutions by applying the mechanisms of selection, crossover and mutation. In our case, each member of the population corresponds to a unique combination of K assets (see also [16] for a similar approach). The initial population is chosen randomly and, in each generation, the top-n members of the population are selected for reproduction. The quality of an individual solution is measured by the value of the Sortino ratio. Previous experiments have indicated that this *elitist* method of selection yields quite good results. After the selection process is completed, a certain percentage of elite members undergo the process of crossover, in which pairs of "parent"-portfolios exchange their constituents, and then the process of mutation, in which certain parts of a portfolio are randomly substituted by other candidate assets.

The second hybrid scheme, implemented in this study, is an ant colony optimisation algorithm, inspired by the foraging behavior of real ant colonies [3]. In nature, ants randomly explore their surrounding environment for food sources. When an individual ant is faced with a potential food source, it carries some of it back to the nest. During the return trip, the ant deposits a certain amount of a chemical substance, called *pheromone*, which helps communicating with the rest of the population. Good-quality food sources are continuously visited by ants, thus enhancing the corresponding pheromone trails. In our setting, each ant represents a "financial" agent whose aim is to select assets for the portfolio. The food sources correspond to available assets and a complete solution consists of a combination of K assets. At first, artificial ants pick their solutions randomly. Then, for a number of generations, each ant updates the solution, based on the *pheromone* values of each asset (food source). Suppose that in some iteration of the algorithm, a portfolio P_k of $k = 1, ..., K-1$ assets is selected. The probability of adding the asset j to the existing portfolio is given by $p_{k,j} = \sum_{i \in P_k} pher_{i,j} / \left(\sum_{i \in P_k} \sum_{h \notin P_k} pher_{i,h} \right)$, where $pher_{i,j}$ is the pheromone value for the "path" connecting asset i with asset j. This ratio essentially compares the total pheromone for the connections between each member of the portfolio and the candidate asset with the sum of pheromone values for the paths connecting each selected asset with any asset not included in the portfolio. At the last step of the algorithm, the pheromone update process takes place, consisting of two parts: firstly, all pheromone values decrease (evaporate) by a certain amount and, secondly, pheromone values of the n-best solutions are reinforced.

4 Experimental Setting

In order to investigate the performance of optimisation metaheuristics, we consider the case where the fund manager's task is to ensemble baskets of Dow Jones Industrial Average (DJIA) stocks. The DJIA index is a popular benchmark of the North American equities market comprised of 30 large and

well-established companies. Our sample data span a period of approximately one trading year (16/11/2010 - 11/11/2011). Daily adjusted closing prices for DJIA member stocks were downloaded from *Yahoo!Finance* service and the associated returns were calculated as $R_t = 1 - P_t/P_{t-1}$, where P_t is the closing price of the trading day t and R_t is the close-to-close daily return. The average portfolio return and the downside standard deviation of returns were computed using the corresponding sample measures over the examined data period. In all optimisation exercises, we assumed that admissible portfolio weights take values between -0.90[1] and 0.90, thus allowing no more than 90% of the initial capital to be invested in a single company. The risk-free rate is 3% per annual, which amounts to an equivalent daily rate of $(1.03^{(1/360)} - 1) \times 360 = 2.96\%$.

Metaheuristic optimisation has only recently started to attract the attention of researchers and practitioners in portfolio management. Keber and Maringer [8] present an empirical application of ant colony systems, genetic algorithms and simulated annealing to the selection of size-constrained portfolios of FTSE 100 stocks. Optimal portfolio allocations were determined using a similar to ours return-to-risk ratio, though, in their case, the risk measure is the ordinary standard deviation of portfolio returns. Maringer [9] provides supplementary evidence on the performance of ant colony systems in selecting cardinality-restricted allocations in three popular stock indices (DAX, FTSE 100 and S&P 100). A review of other research works in metaheuristic portfolio optimisation with an upper limit on the total number of investable assets can be found in [9, 15].

The problem setting assumed in this study differs in several aspects from that reported in other similar applications of metaheuristic techniques. We explicitly introduce floor/ceiling constraints in the formulation of the problem (allowing short-selling of stocks) and also use the downside standard deviation to penalise excess portfolio returns. The latter is a non-quadratic measure of risk which makes the optimisation problem more difficult to solve. What is more, we put more emphasis on exploring the stochastic behaviour of metaheuristic schemes. In the aforementioned research works, authors typically perform several independent runs of each stochastic optimiser and then report the best solution found in all repetitions or some measure of dispersion such as the standard deviation of final scores[2]. This only gives a rough indication of the uncertainty associated with algorithmic performance. In this study, we go deeper into analysing the convergence paths followed by individual algorithms which further allows us to draw conclusions on the optimal execution design.

Cardinality-constrained capital allocations were derived by solving the optimisation problem (1) assuming bundles of $K = \{5, 15, 20\}$ stocks. We made an effort to perform a fair comparison of optimisation techniques, using the

[1] A negative weight means that the corresponding asset is sold short.

[2] An exception is perhaps [9] who also reports intervals within which a certain percent of algorithmic outcomes fall.

same level of computational resources. The population size was set at 100 and all algorithms were terminated after $T = 200$ iterations. To analyse the variability of results, we performed $M = 100$ independent executions from random initial states. For choosing parameter values for each algorithm, we mainly resorted to default values reported in the literature. In fact, we deliberately avoided looking long enough and hard enough for optimal parameter values, as this would bias the results of our study towards specific optimisation techniques. For GA, 10 out of 100 portfolios in each generation were used for reproduction, 95% of the population members were selected for crossover and only a 15% was subject to mutation. For ACO, the number n of elite members was also 10 and the evaporation rate was set to 0.7, thus enforcing quick evaporation of weak pheromone trails.

The ability of metaheuristics to indicate good bundles of assets was evaluated against a simple Monte-Carlo technique. We generated $10,000$ random combinations of K stocks and, for each of them, we used the Levenberg-Marquardt algorithm to solve the subproblem (b), i.e. detecting the capital allocation with the maximum risk-penalised return. If the examined metaheuristics are of some value to the portfolio manager, they should detect asset combinations that are statistically superior to a "blind" search over all possible combinations.

5 Statistical Performance Analysis

Due to space limitations, we only present experimental results for portfolios of size $K = 15$.[3] Results for other cardinalities are available upon request.Amongst all optimisation techniques, GA was the one to detect the portfolio with the maximum Sortino ratio (0.629). This value is taken as the "*global*" optimum (GO) of the problem at hand and all other reported solutions are evaluated against GO.

Our statistical performance analysis is centered around the concept of the *success* or *hit rate*. This is defined as the percentage of algorithmic runs for which the final best solution reported is *at worse* $y\%$ away from GO.[4] This may equivalently be interpreted as an extreme value probability derived from the empirical distribution of algorithmic outcomes (see [6]). The relative frequency of successful runs is, under proper conditions, a consistent estimate of the actual probability of hitting the optimal region in a single run. However, due to the fact that the hybrid optimisation strategies examined in this chapter are computationally demanding, one is confined to a relatively small number of algorithmic restarts to obtain an estimate of this quantity. This has a negative effect on the reliability of the derived values for the success

[3] According to Fig. 1, this is the cardinality value for which the combinatorial complexity reaches its maximum level.

[4] The percentage deviation for a solution y is computed as $1 - y/GO$.

rates. Therefore, it would be advisable to equip our point estimate with a *confidence interval* showing how much the success rate is expected to vary beyond our experimental data. Generally, an analytical formula for the confidence interval is hard to derive, as the exact distribution of the rate is unknown in finite samples. One could rest on asymptotic theory and assume that the hit rate is approximately normally distributed. Despite the fact that the hit rate is a non-negative quantity, extensive simulation experiments show that this normality approximation is rather poor assuming few algorithmic restarts.

In this chapter, we apply *bootstrapping* to estimate the finite-sample distribution of the hit rate statistic. Bootstrapping is a relatively simple method for deriving measures of variability for sample estimates. It relies on resampling to reproduce the sample distribution of the estimated quantity (see e.g. [2] for more information). The bootstrapping procedure applied in this study for calculating confidence intervals on the hit rate is analytically described in the sequel.

Let d_m be a real-valued positive variable showing the percentage deviation from the GO of the best solution reported in the mth run, where $m = 1, ..., M$ and $M = 100$ in our case. The empirical rate $f(x)$ of getting a solution in the optimum region $[(1 - x)GO, GO]$ is defined as $f(x) \equiv (1/M) \sum_{m=1}^{M} 1_{\{d_m \leq x\}}$, where x is the tolerance rate and $1_{\{.\}}$ is the indicator function. We draw $B - 1 = 499$ sets of hypothetically observed deviations $\left\{ d_m^{(b)},\ m = 1, ..., M; b = 1, ..., B \right\}$ by randomly sampling with replacement from the original set of d_m's. For each bootstrapped sample b, we compute the empirical success rate

$$f^{(b)}(x) = (1/M) \sum_{m=1}^{M} 1_{\left\{ d_m^{(b)} \leq x \right\}}$$

and thus form a new sample of success rate points $\left\{ f^{(b)}(x),\ b = 1, ..., B \right\}$.[5] A two-sided confidence interval on the hit rate with overall level of confidence $(1 - \alpha)$ is estimated by calculating the $\alpha/2$- and $(1 - \alpha/2)$-cutoff points from the empirical distribution of $f^{(b)}(x)$'s.

Table 1 reports 95% confidence intervals on the hit rate of each metaheuristic for various levels of deviation from the global optimum solution. For comparison purposes, we report the corresponding rates for the Monte-Carlo portfolio-search method. Note that if we only permit small deviations from the GO (where x is at the order of 1% or 2%), we have practically zero chance of observing, at 95% confidence level, a successful outcome for ACO. GA has a slightly higher probability of reaching the 1%- or 2%-optimum region, although this probability is *not* statistically different from zero at 5% significance. As we increase the width of the optimum region, the success rate quickly peaks up. Particularly poor convergence is observed for Monte-Carlo,

[5] The last observation of this sample is taken equal to $f(x)$, the actual hit rate.

Table 1 Bootstrapped confidence intervals on algorithmic hit rates for a range of tolerance values x.

Deviation (%)	ACO	GA	Monte-Carlo
1	(0.000 0.000)	(0.000 0.041)	(0.000 0.000)
2	(0.000 0.000)	(0.000 0.054)	(0.000 0.000)
3	(0.000 0.036)	(0.000 0.077)	(0.000 0.000)
4	(0.000 0.060)	(0.054 0.169)	(0.000 0.000)
5	(0.033 0.131)	(0.195 0.368)	(0.000 0.000)
8	(0.534 0.716)	(0.937 1.000)	(0.000 0.000)
10	(1.000 1.000)	(1.000 1.000)	(0.000 0.000)
20	(1.000 1.000)	(1.000 1.000)	(0.000 0.001)
30	(1.000 1.000)	(1.000 1.000)	(0.015 0.020)
40	(1.000 1.000)	(1.000 1.000)	(0.186 0.201)

which fails to detect near-optimum allocations with significant probability. Note that if we set x equal to 20%, we are almost 100% sure that a single run of ACO and GO will converge to the optimal region. Still, the chances of getting a near-optimum asset combination with a Monte-Carlo trial is almost zero.

6 Worst Expected Outcome

Recognising the statistical variability of final outcomes, many researchers applying metaheuristic optimisation techniques tend to report deviations from average algorithmic performance. However, in practical applications it might be more useful to focus on the worst algorithmic outcome that is likely to be observed with a certain probability. This *safety-first* approach to evaluating algorithmic performance can be implemented by calculating partial moments and critical points associated with the tails of the distribution of the performance metric of interest.

In Table 2, we report various percentiles of the empirical distribution of solutions attained by each optimisation metaheuristic in 10, 100 or 200 iterations. The numbers presented under the columns "ACO" and "GA" can be interpreted as the maximum *percentage* deviation from the global optimum that is expected with credibility $100\times(1-\alpha)\%$. The last three rows of Table 2 show, for comparison, the α-worst outcome from the Monte-Carlo search[6]. Instead of reporting single estimates, we compute confidence intervals on cutoff points using a bootstrapping procedure similar to that described in Sect. 5. Fig. 2 shows the evolution of the $5th$ percentile over a finer resolution of algorithmic iterations (the grey-shaded region lying between the two neighboring

[6] As Monte Carlo is not an iterative procedure, its results are only comparable with the final outcomes of the metaheuristic optimisation schemes.

Table 2 Confidence intervals on the worst expected deviation of reported solutions for various significance levels and number of generations.

Significance (α)	ACO	GA	Monte-Carlo
Number of generations: 10			
0.10	(16.40 17.20)	(11.33 12.90)	
0.05	(16.70 17.68)	(11.62 13.57)	
0.01	(17.25 18.37)	(12.98 14.18)	
Number of generations: 100			
0.10	(9.78 10.56)	(7.71 8.15)	
0.05	(10.16 11.24)	(7.92 8.55)	
0.01	(10.75 11.63)	(8.24 8.71)	
Number of generations: 200			
0.10	(8.64 9.15)	(7.29 7.87)	(55.95 56.44)
0.05	(8.90 9.49)	(7.62 8.26)	(58.65 59.37)
0.01	(9.23 9.92)	(7.91 8.71)	(63.89 65.12)

curves corresponds to a 95% confidence region on the estimated cutoff point).

As observed from Table 2 and Fig. 2, if algorithms are allowed to iterate only 10 times, the maximum deviation from the global optimum can be at the

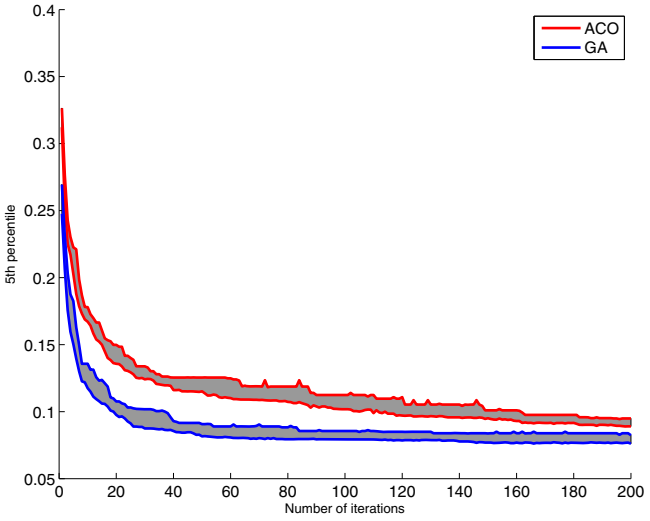


Fig. 2 The evolution of the worst expected outcome (as measured by the 5th percentile of the empirical distribution of solutions) with the number of iterations.

order of 17% for ACO and 12% for GA. Provided that we want to be 99% confident that we have actually witnessed the worst algorithmic outcome, these numbers climb up to 18.37% for ACO and 14.18% for GA. Note that at all significance levels, the right-hand side of the GA confidence intervals lie below the left-hand side of the ACO confidence intervals, meaning that GA is *statistically superior* to ACO in terms of the worst expected outcome. The relative ranking is also preserved even after more iterations have elapsed, although maximum expected deviations in these cases are much lower. As seen from Fig. 2, by running at least 20 iterations from GA and 160 from ACO, the researcher can be quite confident that the Sortino ratio of the reported optimal portfolio will not under-perform the best-ever solution (0.629) by more than roughly 10%. Whether this deviation is acceptable or not has to do with the portfolio manager's aspiration levels. The point, however, is that with the analysis presented we are able to quantify the risk of severe algorithmic under-performance and, eventually, know what to expect from each optimisation scheme.

7 Execution Design

An important issue in metaheuristic optimisation is the choice of values for algorithmic parameters, especially the optimal trade-off between the exploration and exploitation of the solution space. One the one extreme, the researcher may fix a relatively small number of parallel agents and spend more computational resources on locating the optimum region with higher accuracy. On the other edge, he/she might choose to cutdown on the number of generations with the purpose of increasing the population size and thus performing a more consistent exploration of the solution space. In this section, we show how the statistical dispersion analysis presented in previous sections could provide some guidance in this dilemma.

Assuming that enough independent runs can be executed for each algorithm, one might ask for the minimal number of restarts such that at least one successful outcome is observed with high confidence. Based on the theory of Binomial distributions, we can design a procedure for calculating the required algorithmic repetitions [6] (see also [14] for a detailed discussion). The idea is to see each run of the optimisation technique as a Bernoulli trial with probability of success equal to $f(y,t)$. Extending our previous notation, this is the relative frequency by which the algorithm has detected in $t = 1, ..., 200$ iterations a solution that is at worse $(100 \times y)\%$ inferior to the global optimum portfolio. Fixing the value of t, we can estimate the minimal number of trials such that the probability of observing *at least one* successful outcome is no less than a certain threshold. The desired quantity can be computed by inverting the Binomial cumulative probability function.

Table 3 Stochastic convergence analysis and optimal operation of metaheuristic optimisation techniques. Confidence intervals are shown in parentheses.

Number of iterations	ACO		GA	
	Hit rate (5%)	Algorithmic restarts	Hit rate (5%)	Algorithmic restarts
1	(0.000, 0.000)	($>10^4$, $>10^4$)	(0.000, 0.000)	($>10^4$, $>10^4$)
10	(0.000, 0.000)	($>10^4$, $>10^4$)	(0.014, 0.096)	(30, 211)
20	(0.000, 0.046)	(64, $>10^4$)	(0.029, 0.125)	(23, 101)
50	(0.000, 0.046)	(65, $>10^4$)	(0.075, 0.198)	(14, 39)
100	(0.000, 0.076)	(39, $>10^4$)	(0.138, 0.305)	(9, 21)
150	(0.014, 0.093)	(31, 208)	(0.168, 0.335)	(8, 17)
170	(0.031, 0.122)	(24, 97)	(0.188, 0.350)	(7, 15)
200	(0.033, 0.131)	(22, 96)	(0.195, 0.368)	(7, 14)

Table 3 reports, for each optimisation technique, interval estimates on the probability of observing a solution in the 5%-optimal region as well as on the number of repetitions required to detect the optimum region with 95% confidence.[7] Note that in a few cases (particularly when the number of iterations is quite small), the probability of a successful outcome is almost zero (to three significant digits). This means that the algorithm has to be redeployed quite many times (greater than $10,000$) on the same problem instance, so that the portfolio manager can be confident that optimality has been achieved. In the case of ACO, the empirical success rate is statistically insignificant even after a considerable number of 100 iterations have elapsed. On the contrary, for GA, the hit rate peaks up faster with the number of generations, thus implying fewer algorithmic restarts. For example, with 211 initialisations from random states, the GA is 95% likely to detect close-to-optimal solutions in only 10 iterations. If the algorithm is allowed to iterate more times, optimality can be attained in less than 20 or 15 generations.

One can take the presented analysis one step further and derive the combination of restarts/iterations guaranteeing that the minimum amount of computational resources is spent on the particular optimisation exercise. This optimal trade-off can be easily computed for each metaheuristic by multiplying the right-hand side of the interval estimate of required independent runs with the corresponding number of iterations shown at the first column of Table 3. For the subset of values reported here, we can easily infer that ACO is more efficiently executed by deploying 170 iterations from 97 random initial states. The corresponding optimal trade-off for GA is 50 iterations and 39 independent runs.

[7] Results for values of the tolerance rate other than 5% are available from the authors upon request. Monte Carlo experiments are omitted from this Table, as the probability of converging to a solution that is at most 5% away from the *GO* has not been found statistically different from zero (see Table 1).

8 Conclusions and Further Research

This chapter performed an empirical analysis of the convergence properties of two hybrid computational schemes, based on genetic algorithms and ant colonies, in the framework of cardinality-constrained portfolio optimisation. We report several indicators of algorithmic performance picturing different aspects of the statistical variability of outcomes. The information conveyed by these performance metrics can be further utilised to derive the optimal trade-off between the number of generations and restarts. This figure is very important when it comes to deciding how to allocate the available computational resources given the complexity of the optimisation problem. We finally proposed a technique for deriving confidence intervals on various sample estimates, showing how the value of each performance metric is expected to vary on unseen data. Both metaheuristics examined in this study have been proven quite effective in handling the complexities of the optimisation problem at hand. The subgroups of stocks suggested by GA or ACO consistently deliver higher Sortino ratios than those reported by a Monte-Carlo search over the space of feasible asset combinations.

The techniques presented in this study are general-purpose and can be applied to a wider range of stochastic metaheuristics or optimisation problems (see e.g. [6] for a discussion). However, at the current stage of development, they mainly serve as an *ex-post* evaluation tool for measuring the convergence rate of a metaheuristic or performing pairwise algorithmic comparisons on a particular problem instance. The extent to which the results from such an analysis provide useful information on how to operate each algorithm on a similar problem setting is still an issue under investigation.

References

1. Barrero DF, Castaño B, R-Moreno MD, Camacho D (2011) Statistical Distribution of Generation-to-Success in GP: Application to Model Accumulated Success Probability. *Proc. 14th European Conf. on Genetic Programming (EuroGP 2011)*, 154–165
2. Chernick MR (1999) *Bootstrap Methods: A practitioner's guide*. Wiley Series in Probability and Statistics
3. Dorigo M, Stützle M (2004) *Ant Colony Optimization*. MIT Press
4. Dreo J, Petrowski A, Siarry P, Taillard E (2006) *Metaheuristics for Hard Optimization: Methods and Case Studies*. Springer, Berlin Heidelberg New York
5. García S, Molina D, Lozano M, Herrera F (2009) A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'2005 Special Session on Real Parameter Optimization. *Journal of Heuristics* 15:617–644
6. Gilli M, Winker P (2009) Heuristic Optimization Methods in Econometrics. In: Belsley D, Kontoghiorghes E (eds.) *Handbook of Computational Econometrics*, 81–119. J. Wiley & Sons, Chichester
7. Holland JH (1992) Genetic Algorithms. *Scientific American* 267:66–72

8. Keber C, Maringer D (2001) On Genes, Insects and Crystals: Determing Marginal Diversification Effects with Nature Based Methods. *Computing in Economics and Finance* 152. Society for Computational Economics
9. Maringer D (2006) Small is Beautiful: Diversification with a Limited Number of Assets. *CCFEA Working Paper Series (WP005-06)*. University of Essex
10. Moré JJ (1978) The Levenberg-Marquardt algorithm: Implementation and theory. *Lecture Notes in Mathematics* 630:104–116. Springer, Berlin
11. Sortino AF, Price NL (1994) Performance Measurement in a Downside Risk Framework. *The Journal of Investing* 64:59–64
12. Talbi E-G (2009) *Metaheuristics: From Design to Implementation*. J. Wiley & Sons, Chichester
13. Thomaidis NS (2010) Active Portfolio Management from a Fuzzy Multi-objective Programming Perspective. In: Cecilia Di Chio *et al.* (eds) *Lecture Notes in Computer Science* 6025:222–231. Springer, Berlin
14. Thomaidis NS (2011) A soft computing approach to enhanced indexation. In: Brabazon A, O'Neill M, Maringer D (eds) *Natural Computing in Computational Finance (Volume IV)*. Springer, Berlin
15. Vassiliadis V, Thomaidis NS, Dounias G (2009) Active Portfolio Management under a Downside Risk Framework: Comparison of a Hybrid Nature — Inspired Scheme. In: Corchado E. *et al.* (eds.) *HAIS 2009. Lecture Notes in Computer Science* 5572:702–712. Springer, Berlin
16. Vassiliadis V, Thomaidis NS, Dounias G (2011) On the Performance and Convergence Properties of Hybrid Intelligent Schemes: Application on Portfolio Optimization Domain. In: Di Chio C *et al.* (eds.) *EvoApplications 2011. Lecture Notes in Computer Science* 6625:131–140. Springer, Berlin
17. Weise T (2009) *Global Optimization Algorithms: Theory and Application*. E-book available from `http://www.it-weise.de/`

Comparison of Multi-objective Algorithms Applied to Feature Selection

Özlem Türkşen[1], Susana M. Vieira[2], José F.A. Madeira[2,3], Ayşen Apaydın[1], and João M.C. Sousa[2]

Abstract The feature selection problem can be formulated as a multi-objective optimization (MOO) problem, as it involves the minimization of the feature subset cardinality and the misclassification error. In this chapter, a comparison of MOO algorithms applied to feature selection is presented. The used MOO methods are: Nondominated Sorting Genetic Algorithm II (NSGA-II), Archived Multi Objective Simulated Annealing (AMOSA), and Direct Multi Search (DMS). To test the feature subset solutions, Takagi-Sugeno fuzzy models are used as classifiers. To solve the feature selection problem, AMOSA was adapted to deal with discrete optimization. The multi-objective methods are applied to four benchmark datasets used in the literature and the obtained results are compared and discussed.

1 Introduction

Generally, real-world data sets tend to be complex, very large, and normally contain many irrelevant features. One of the most important steps in data analysis for classification is feature selection, which has been an active research area on many fields, such as data mining, pattern recognition, image understanding, machine learning or statistics. The main idea of feature selection is to choose a subset of the available features, by eliminating redundant features with little or no predictive information. There are two key decisions involved in the feature subset selection problem: (i) the number of selected

[1] Ankara University, Faculty of Science, Statistics Department, 06100, Ankara, Turkey, {Ozlem.Turksen,Aysen.Apaydin}@science.ankara.edu.tr
[2] Technical University of Lisbon, Instituto Superior Técnico, Dept. of Mechanical Engineering, CIS - IDMEC/LAETA, Lisbon, Portugal, {susana.vieira,jmsousa}@ist.utl.pt
[3] ISEL, Rua Conselheiro Emidio Navarro, 1, 1959-007 Lisboa, Portugal, jaguilar@dem.ist.utl.pt

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 359–375.
springerlink.com

features and (ii) the best features to be selected [28]. An effective feature selection method can minimize the classification error, improve the prediction accuracy, and also discover the relevant features. From this point of view, it is possible to say that feature subset selection is a multi-objective optimization problem. Recently, the multi-objective feature selection (MOFS) problem was addressed in many studies. In [14], the feature selection problem is seen as a multi-objective problem and the Niched Pareto Genetic Algorithm (NPGA), which uses a commonality-based crossover operator, is applied to solve the MOFS problem, using neural models as classifiers. A variation of NPGA with a one-nearest neighbor classifier is applied to the MOFS problem in [13]. In [15], the application of the Multi Objective Genetic Algorithm (MOGA) is proposed for feature subset selection on a number of neural and fuzzy models together with fast subset evaluation techniques. Further, MOGA was used with different classifiers, namely, fuzzy rule based classification [8], back-propagation neural networks [19], and support vector machines [5], to solve the MOFS problem. There are some studies on the use of the nondominated sorting genetic algorithm (NSGA), firstly proposed in [26], for MOFS with different wrapper methods, such as neural networks [23] and decision trees [32], on different data sets. In [17, 22, 30, 12, 18, 24], the nondominated sorting genetic algorithm II (NSGA-II), one of the most efficient multi-objective algorithms, was used to solve the MOFS problem using different classification methods.

Archived multi-objective simulated annealing (AMOSA), which was proposed in [3], is an efficient multi-objective version of the simulated annealing algorithm, based on Pareto dominance. AMOSA incorporates a novel concept of the amount of dominance, in order to determine the acceptance of a new solution, as in NSGA-II. AMOSA was mainly used for continuous multi-objective problems, except in [31] where it was applied to gene selection. In this chapter, AMOSA was adapted for the feature selection problem, and it is called Modified AMOSA.

Direct multisearch (DMS) is a novel MOO algorithm proposed in [9]. This method is inspired by the search/poll paradigm of direct-search methods of the directional type and uses the concept of Pareto dominance to maintain a list of non-dominated points (from which the new iterates or poll centers are chosen). The aim of this method is to generate as many points in the Pareto front as possible from the polling procedure itself, while keeping the whole framework general enough to accommodate other disseminating strategies. This chapter presents a comparison of derivative-free multi-objective algorithms, which are NSGA-II, Modified AMOSA and DMS, for the feature selection problem with two different objectives: minimizing the number of features and minimizing the misclassification rate. The chapter is organized as follows: the next section presents a multi-objective formulation of the feature selection problem and a brief description of fuzzy modeling for classification. The derivative-free multi-objective algorithms for MOFS, namely NSGA-II, Modified AMOSA and DMS, are presented in Section 3. In Section 4, the

results obtained for several benchmark databases are presented, and a comparison of the studied multi-objective algorithms is made. Some conclusions are drawn in Section 5 and possible future work is discussed.

2 Feature Selection

Feature selection is the process of selecting a subset of the available features to use in empirical modeling. A fundamental problem of feature selection is to determine a minimal subset of n features from the complete set of the features $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, with $n < N$, without sacrificing accuracy. This means that there are 2^N possible feature subsets, which makes a brute-force approach (enumerating and testing all feature subsets) infeasible in most cases. It can be said that the main goal of feature selection is to reduce the number of features used in classification while maintaining an acceptable classification accuracy. From this perspective, a feature selection problem can be seen as a multi-objective problem with two objectives: minimization of the number of features and of the error rate of the classifier. In this chapter, a fuzzy classifier is built for each feature subset to evaluate the classification error. The solutions are evaluated by using fuzzy models, as they are universal approximators and can be interpretable under certain conditions.

2.1 Feature Selection as a Multi-objective Optimization Problem

A multi-objective optimization problem can be mathematically formulated as (see [21] for a more complete treatment):

$$\begin{aligned} &\min\ F(\mathbf{x}) \equiv (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))^T \\ &\textit{subject to} \qquad \mathbf{x} \in \Omega \end{aligned} \tag{1}$$

where x is the vector of decisions or design variables belonging to the feasible region $\Omega \subseteq \mathbb{R}^n$ and $F(\mathbf{x}) \in \mathbb{R}^m$ is a vector of m objective functions. The word "min" in (1) means that we want to minimize all objective functions simultaneously. Note that this is a general formulation, which exploits that maximizing an objective function f_j is equivalent to minimizing $-f_j$.

The solution of the problem given in (1) is a set of solutions that are called Pareto optimal in the general framework of multi-objective optimization. A solution is said to be Pareto optimal if it is not dominated by any other solution available in the search space Ω (see [21] for a more complete treatment).

In feature selection it is common to encode a feature subset as a binary vector $\mathbf{x} \in \{0,1\}^N$, where N is the number of available features. This vector states for each feature whether it is selected ($x_i = 1$) or not ($x_i = 0$), see Fig. 1. The MOO algorithm for the feature selection problem consists of finding the set of features that simultaneously minimize two objectives: $f_1(\mathbf{x}) = \sum_{k=1}^{N} \mathbf{x}_k$ which is the number of selected features and $f_2(\mathbf{x}) = (1 - accuracy(\mathbf{x}))$ which is equivalent to maximize the accuracy.

2.2 Fuzzy Modeling for Feature Selection

Fuzzy models are suitable to deal with vague, imprecise and uncertain knowledge and data. These models use rules and logical connectives to establish relations between the features defined to derive the model. Three general methods for fuzzy classifier design can be distinguished [4, 25]: the regression method, the discriminant method and the maximum compatibility method. In the discriminant method, the classification is based on the largest discriminant function, which is associated with a certain class, regardless of the values or definitions of other discriminant functions. Hence, the classification decision does not change if the discriminant functions are transformed monotonically. Since this is a useful property, the discriminant method is used in this work for classification. The discriminant functions can be implemented as fuzzy inference systems, which can be Takagi-Sugeno (TS) fuzzy models [27]. When TS fuzzy systems are used, each discriminant function consists of rules of the type

$$\begin{aligned} \text{Rule } R_i^c : \text{If } x_1 \text{ is } A_{i1}^c \text{ and } \ldots \text{ and } x_n \text{ is } A_{in}^c \\ \text{then } d_i^c(\mathbf{x}) = f_i^c(\mathbf{x}),\ i = 1, 2, \ldots, K, \end{aligned}$$

where c is the number of classes, K is the number of fuzzy rules, and f_i^c is the consequent function for rule R_i^c. Please note that the antecedent parts of the rules can be different for different discriminants, as well as the consequents. Therefore, the output of each discriminant function $d_c(\mathbf{x})$ can be interpreted as a score (or evidence) for the associated class c given the input feature vector $\mathbf{x}_n$. The discriminant function for class c, with $c = 1, \ldots, C$ is computed by aggregating the contributions of the individual rules:

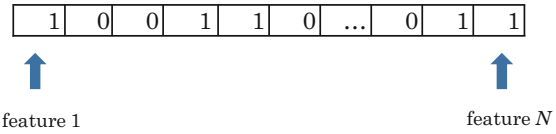


Fig. 1 Binary vector with N components.

$$d_c(\mathbf{x}) = \frac{\sum_{i=1}^{K} \beta_i f_i^c(\mathbf{x})}{\sum_{i=1}^{K} \beta_i}. \quad (2)$$

where $\beta_i = \prod_{j=1}^{n} \mu_{A_{ij}^c}(\mathbf{x})$ is the degree of activation of rule i of class c and $\mu_{A_{ij}^c}(\mathbf{x}) : \mathbb{R} \rightarrow [0, 1]$.

The input data consists of tuples $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}, y_i]^T$with $i \in \{1, \ldots, N\}$, where the x_{ij}, $j \in 1, ...M$, are the values of the features and y_i is the class of the ith case. From this data and the number of rules K, the antecedent fuzzy sets A_{ij}, and the consequent parameters are determined by means of fuzzy clustering [25]. Note that the class is used as an input to the clustering algorithm. This paper uses the Gustafson-Kessel (GK) [16] clustering algorithm to compute the fuzzy partition matrix. Each identified cluster provides a local characteristic behavior of the system and each cluster defines a rule. The consequent parameters for each rule are obtained as a weighted ordinary least-square estimate.

3 Multi-objective Algorithms for Feature Selection

Derivative-free multi-objective algorithms can be used for feature selection problems, as most of these algorithms can deal with a set of possible solutions simultaneously. This means that several members of the Pareto optimal set can be found in a single run without any assumptions on continuity and differentiability of functions [23]. A comprehensive review of multi-objective algorithms can be found in [6, 10]. In this chapter, NSGA-II, Modified AMOSA, and DMS, which are described in the following sections, are used for the optimization of the MOFS problem.

3.1 NSGA-II for Multi-objective Feature Selection

NSGA-II has been successfully applied in many applications, such as image processing, bioinformatics, etc. [11]. The main characteristic of this algorithm is to use the fast non-dominated sorting technique and a crowding distance to construct population fronts that dominate each other in a domination rank. To implement NSGA-II in the MOFS problem, first a random population representing different points in the search space is created with the size n_{pop}, as in [18]. Each chromosome $C_i, i = 1, 2, \ldots, n_{pop}$, is binary, and the encoding of a chromosome was represented in Fig. 1, where a chromosome has n bits equal to "1". These features are the ones used to construct the fuzzy classifier. The search process of NSGA-II continues until the number of generations n_{gen} is reached. Tournament selection is used based on two criteria: rank and crowding distance, with rank taking precedence, see [11]. NSGA-II also

incorporates an elitism scheme to maintain the best solutions; individuals with higher crowding distances have higher fitness values. NSGA-II for MOFS uses the single-point crossover method, which randomly chooses a crossing site along the string and exchanges all bits on the right side of the crossing site [10]. The mutation operator used in this work is the uniform mutation operator [20], which operates on each bit separately and randomly changes the bit's. If the chromosome is filled with "0"s, i.e., the feature subset is empty, a gene of the chromosome is selected randomly and replaced by "1" to obtain a non-empty feature subset.

3.2 Modified AMOSA for Multi-objective Feature Selection

AMOSA is a generalized version of simulated annealing for multi-objective optimization problems, which was proposed in [3]. AMOSA, a Pareto dominance based simulated annealing method, incorporates the concept of an *Archive* where the nondominated solutions seen so far are stored. In contrast to the original suggestion [3], we use a fixed-sized *Archive*, AL, instead of a soft and a hard limit. Similarly to the original AMOSA, initially a random *Archive* is generated with $\beta \times AL$ solutions, where β is a constant, $0 < \beta \leq 3$. As in NSGA-II, the solutions are encoded as a binary vector. The fitness evaluation is done for each *Archive* solution. The solutions in the *Archive* are sorted by using a domination relation. Only the obtained nondominated solutions are used to initialize the *Archive*. Here, the fast nondominated sorting mechanism is used to rank the solutions [11]. During the search, if the number of solutions in the *Archive* is higher than AL, the first AL solutions in the rank order are chosen from the *Archive*. Eventually, the ranked "1" solutions will constitute the Pareto front. Table 1 describes the parameters that need to be set *a priori*.

The search is started with a solution that is chosen randomly from the solutions in the *Archive*. This is taken as the current feature subset $F_{current}$, or

Table 1 Definition of the parameters for Modified AMOSA

Parameter	Description
Archive	Set of nondominated solutions
AL	Limit size of the Archive
β	Constant for the initial size of the Archive
T_{max}	Maximum (initial) temperature
T_{min}	Minimum (final) temperature
α	Cooling rate in simulated annealing
iter	Number of iterations at each temperature

Algorithm 1 Modified AMOSA for feature selection

```
Set AL, β, T_max, T_min, α, iter and T = T_max.
Initialize the Archive.
Choose randomly a subset of features from the nondominated solutions in the Archive
and set it as F_current
Compute the probability of being accepted P_current using (4).
while T > T_min do
  for k = 0 to iter do
    Create a new subset of features F_new, using the perturbation scheme
    Compute P_new
    Check the domination status of F_new with respect to the F_current and the present
    solutions in the Archive
    Select the next F_current using the domination status of the original AMOSA [3].
  end for
  update T = α × T
  if | Archive | ≥ AL then
    choose the first AL solutions from the Archive.
  end if
end while
```

the initial solution, at temperature $T = T_{max}$. To create a new solution, the $F_{current}$ solution is perturbed and a new feature subset F_{new} is generated and the solution is evaluated using the objective function. The perturbation is computed using a Laplace distribution to determine which decision variables should be mutated. Thereafter, the domination status of F_{new} is checked with respect to $F_{current}$ and to the solutions in the *Archive*. The nondominated solutions are defined using the acceptance concept from the original AMOSA [3]. Given two solutions a and b, where a dominates b, the amount of domination is defined as follows:

$$\Delta dom_{a,b} = \prod_{i=1, f_i(a) \neq f_i(b)}^{m} \left(\frac{\mid f_i(a) - f_i(b) \mid}{R_i} \right) \tag{3}$$

where m is the number of objectives, $f_i(a)$ and $f_i(b)$ are the ith objective values for the two different solutions and R_i is the range of the objective function. R_i is determined using the solutions in the *Archive*, in the current and in the new feature subsets. Based on the domination status, a number of cases can arise: (i) accept F_{new}, (ii) accept $F_{current}$ or (iii) accept a solution from the *Archive*. The acceptance is calculated based on the applicable case. The *Archive* limit is maintained and the content is continuously updated during the search. Whenever an unfavorable move is considered for acceptance, the probability of acceptance is calculated as:

$$P = \frac{1}{1 + \exp(\Delta dom \times T)} \tag{4}$$

where T is the temperature, and Δdom is calculated as in (3) [31]. The process is repeated *iter* times for each temperature, which is annealed with a cooling rate of $\alpha < 1$ until the minimum temperature T_{min} is reached. The process then stops, and the *Archive* contains the nondominated solutions. The steps of the Modified AMOSA applied to the feature selection problem are presented in Algorithm 1.

3.3 DMS for Multi-objective Feature Selection

Direct MultiSearch (DMS) is a novel derivative-free method for multi-objective optimization, which does not aggregate any of the objective functions [9]. DMS extends to MOO all types of direct-search methods that are of a directional type such as pattern search and generalized pattern search, generating set search, and mesh adaptive direct search [7]. This approach is called direct multisearch since it naturally generalizes direct search (of directional type) from single to multi-objective optimization.

The principles of DMS are extremely simple. Instead of updating a single point per iteration, it updates a list of feasible nondominated points. For a more detailed description of the algorithm please see [9]. The original DMS was developed for real-valued variables in the search space. This algorithm was adapted to cope with the discrete feature selection optimization problem. Each DMS solution, $\mathbf{x}_{DMS} = (x_1, \ldots, x_N) \in \mathbb{R}^N$ is converted into a binary solution, $\mathbf{x} = (x_1, \ldots, x_N), \forall i = 1, \ldots, N : x_i \in \{0, 1\}$, using the threshold

$$\delta = 0.5, \text{ i.e., } x_i = \begin{cases} 1 \ if \ x_i \geq 0.5 \\ 0 \ if \ x_i < 0.5 \end{cases} \tag{5}$$

4 Experimental Results

The derivative-free multi-objective algorithms for feature selection are applied to data sets taken from some well known benchmarks in the UCI Machine Learning Repository [2].

4.1 Data Sets

Wisconsin breast cancer original (WBCO), Wisconsin diagnostic breast cancer (WDBC), Wisconsin prognostic breast cancer (WPBC) and Sonar, were used to test the NSGA-II, Modified AMOSA, and DMS algorithms. Table 2 summarizes some general information regarding these datasets.

Table 2 Description of the used data sets.

No	data sets	# features	Classes	# samples
1	WBCO	9 (integer)	2	699
2	WDBC	32 (real)	2	569
3	WPBC	34 (real)	2	198
4	Sonar	60 (real & integer)	2	208

The MOO algorithms were implemented in Matlab. The parameter settings of NSGA-II and modified AMOSA used in the experiments are presented in Table 3 and Table 4, respectively.

Table 3 NSGA-II parameter values used in the experiments.

Data set	n_{pop}	P_{cross}	P_{mut}	n_{gen}
WBCO	100	0.8	1/9	100
WDBC	100	0.8	1/30	200
WPBC	100	0.8	1/32	200
Sonar	100	0.8	1/60	500

Table 4 Modified AMOSA parameter values used in the experiments.

Data set	AL	β	T_{max}	T_{min}	α	$iter$
WBCO	100	1.5	200	0.001	0.81	50
WDBC	100	1.5	200	0.001	0.81	400
WPBC	100	1.5	200	0.001	0.81	400
Sonar	100	1.5	200	0.001	0.81	500

The default parameters of DMS were used (version 0.2, May 2011) without cache. This DMS version is freely available for research, educational or commercial use, under a GNU lesser general public license [1].

In this chapter, 10-fold cross validation is used. After the application of multi-objective algorithms and selection of the features, the fuzzy classification models are validated using the test subsets. The prediction performance of the classifier is estimated by considering the average classification accuracy of the 10-fold cross validation experiments.

4.2 Results for Different Data Sets

The NSGA-II, Modified AMOSA, and DMS methods are applied to minimize both the size of feature subsets and the average misclassification rates for all data sets. The number of fuzzy rules is equal to 2 for WBCO and 3 for the other databases used in this work.

4.2.1 Wisconsin Breast Cancer Original

The WBCO data is widely used to test the effectiveness of classification algorithms. The aim of the classification is to distinguish between benign and malignant cancers based on nine measurements (attributes): clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. The attributes have integer values in the range $[1, 10]$. The original database contains 699 instances. However 16 of these are excluded as they are incomplete, which is common in data mining. The class distribution is 65.5% benign and 34.5% malignant [29].

Table 5 Feature subsets for WBCO data set with 10 fold cross-validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	$\{2\}$	$\{2\}$	$\{2\}$	7.153
2	$\{2, 6\}$	$\{2, 6\}$	$\{2, 6\}$	4.721
3	$\{1, 2, 6\}$	$\{1, 2, 6\}$	-	4.435
3	-	-	$\{1, 3, 6\}$	3.720
4	-	-	$\{1, 3, 4, 6\}$	3.577
5	-	-	$\{1, 2, 3, 5, 6\}$	3.434

Table 5 shows the set of nondominated solutions obtained by NSGA-II, Modified AMOSA, and DMS. The obtained feature subsets are similar for the three algorithms.

In Figure 2, the nondominated solutions for each algorithm are presented, and it is shown that DMS presents the best results.

4.2.2 Wisconsin Diagnostic Breast Cancer

In WDBC, features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image. The attribute information for WDBC is as follows: ID numbers, diagnosis (M = malignant, B = benign) and ten real-valued features are com-

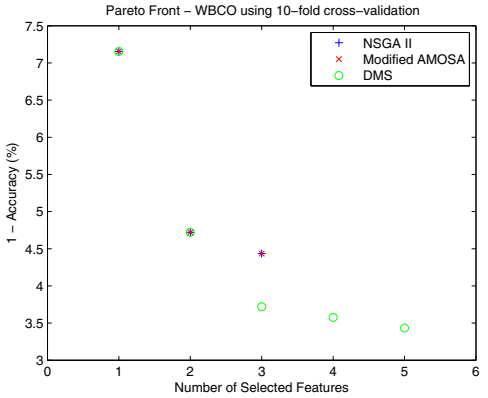


Fig. 2 Comparison of algorithms for WBCO data set.

puted for each cell: nucleus radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness, concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry and fractal dimension [2].

Table 6 Feature subsets for WDBC data set with 10 fold cross validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	{24}	-	{24}	8.260
2	{24, 28}	-	-	6.151
2	-	-	{21, 25}	4.745
3	{22, 24, 28}	-	-	4.569
3	-	-	{2, 21, 25}	4.042
3	-	-	{2, 24, 28}	4.042
3	-	-	{22, 24, 29}	4.042
3	-	-	{24, 25, 29}	4.042
4	{8, 22, 24, 25}	-	-	4.394
4	-	{21, 26, 29, 30}	-	6.503
4	-	-	{2, 3, 24, 25}	3.339
4	-	-	{2, 21, 28, 29}	3.339
5	{8, 10, 22, 24, 25}	-	-	3.866
5	-	{2, 3, 8, 21, 29}	-	5.800
5	-	-	{2, 14, 21, 25, 28}	2.812
5	-	-	{14, 21, 22, 25, 28}	2.812
6	-	-	{2, 14, 21, 25, 28, 29}	2.636
6	-	-	{2, 14, 24, 25, 28, 29}	2.636

The obtained nondominated solutions using NSGA-II, Modified AMOSA and DMS are summarized in Table 6. Figure 3 presents the average misclassification rates and feature subset cardinality of the three algorithms. DMS is clearly the best, followed by NSGA-II and Modified AMOSA, which yields quite poor results when compared to the other two algorithms.

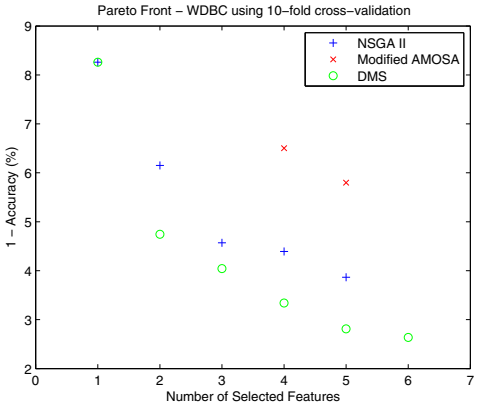


Fig. 3 Comparison of algorithms for WDBC data set.

4.2.3 Wisconsin Prognostic Breast Cancer

In this dataset, each record represents follow-up data for one breast cancer case. These are consecutive patients and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The first 30 features of the WPBC data set are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image [2]. Table 7 shows the feature subsets obtained by DMS, NSGA-II and Modified AMOSA. Figure 4 presents the results obtained by the three algorithms. DMS is the best of the three algorithms, and can find much more points on the Pareto front. In this case, NSGA-II and Modified AMOSA present similar results, both yielding a small number of solutions.

4.2.4 Sonar Data Set

The sonar data set contains information of 208 objects and 60 attributes. The objects are classified in two classes: “rock” and “mine”. A data frame with 208 observations and 61 variables is used. The first 60 represent the energy

Table 7 Feature subsets for WPBC data set with 10 fold cross validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	{25}	-	-	23.74
1	-	-	{5}	22.73
2	{1, 25}	-	{1, 25}	19.70
2	-	-	{1, 22}	19.70
3	-	{11, 23, 24}	-	21.72
3	-	-	{1, 13, 25}	18.18
4	-	{1, 3, 7, 22}	-	18.69
4	-	-	{1, 13, 22, 32}	17.17
5	{1, 6, 8, 13, 25}	-	-	19.19
5	-	-	{1, 13, 20, 22, 32}	16.67
5	-	-	{1, 13, 24, 26, 32}	16.67
6	{1, 6, 8, 13, 19, 25}	-	-	18.18
6	-	-	{1, 6, 11, 13, 18, 32}	16.16
6	-	-	{1, 13, 22, 26, 27, 32}	16.16
7	-	-	{1, 13, 20, 22, 26, 27, 32}	15.66
10	-	{1, 2, 5, 8, 13, 14, 15, 17, 22, 24}	-	18.18
12	-	-	{1, 2, 6, 11, 12, 13, 14, 17, 18, 22, 24, 32}	14.65
14	-	-	{1, 2, 7, 9, 12, 13, 14, 17, 18, 20, 22, 24, 26, 29}	13.64
20	-	-	{1, 2, 5, 9, 12, 13, 14, 16, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 31}	13.13

within a particular frequency band, integrated over a certain period of time. The last column contains the class labels. There are two classes "0" if the object is a rock, and "1" if the object is a mine (metal cylinder) [2].

This data set is an interesting challenge for the proposed algorithm as the number of features is bigger than the usual benchmark examples. The obtained feature subsets are given in Table 8. Note that Modified AMOSA cannot find models with less than 13 features. On the other hand, NSGA-II has only results up to 7 features.

Figure 5 show the results obtained by the three algorithms. DMS presents clearly the best results, also for a wider spread of number of features. NSGA-II has good results for a small number of features. The results using Modified AMOSA are far from the Pareto front.

4.2.5 Discussion

By analyzing and comparing the four datasets, it can be concluded that Modified AMOSA was not able to deal with the feature cardinality in the

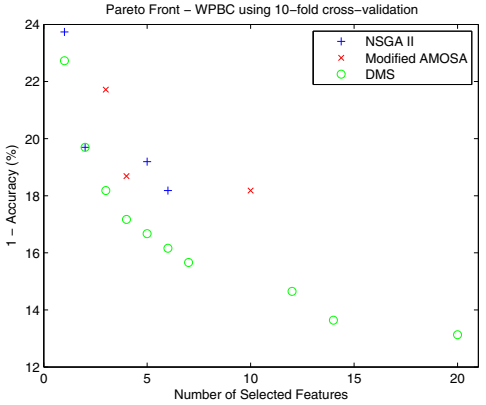


Fig. 4 Comparison of algorithms for WPBC data set.

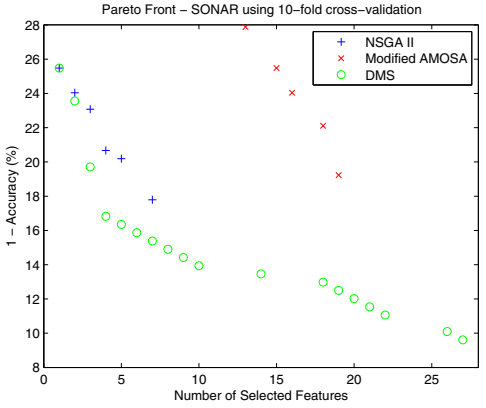


Fig. 5 Comparison of algorithms for the sonar data set.

objective function. NSGA-II proved to be a good algorithm for MOO, as expected. However, the very recent DMS algorithm is clearly the best of the three tested algorithms.

5 Conclusions

In this chapter, the feature selection problem is approached as a multi-objective problem. Two of the most important objectives for the feature selection problem were addressed: the minimization of the feature subset cardinality and the minimization of the classification error.

Table 8 Feature subsets for sonar data set with 10 fold cross validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	{11}	-	{11}	25.48
2	{11, 16}	-	-	24.04
2	-	-	{11, 17}	23.56
3	{11, 16, 21}	-	-	23.08
3	-	-	{11, 18, 19}	19.71
4	{11, 15, 21, 46}	-	-	20.67
4	-	-	{11, 17, 19, 52}	16.83
5	{11, 15, 21, 38, 46}	-	-	20.19
5	-	-	{11, 17, 36, 45, 54}	16.35
6	-	-	{3, 9, 11, 18, 19, 54}	15.87
7	{3, 4, 11, 14, 21, 36, 46}	-	-	17.79
7	-	-	{3, 9, 11, 18, 19, 54, 60}	15.38
8	-	-	{5, 11, 17, 18, 36, 39, 46, 59}	14.90
9	-	-	{5, 11, 14, 17, 18, 36, 39, 46, 59}	14.42
10	-	-	{5, 11, 14, 17, 18, 36, 39, 41, 46, 59}	13.94
13	-	{1, 4, 7, 12, 13, 15, 16, 17, 23, 27, 36, 49, 59}	-	27.88
14	-	-	{7, 11, 17, 18, 23, 25, 30, 31, 35, 36, 44, 49, 50, 52}	13.46
15	-	{3, 4, 11, 12, 13, 15, 17, 29, 30, 34, 35, 40, 49, 59, 60}	-	25.48
16	-	{1, 3, 4, 8, 9, 12, 13, 15, 24, 25, 27, 33, 35, 36, 40, 47}	-	24.04
18	-	{3, 11, 12, 13, 17, 21, 30, 34, 35, 36, 40, 45, 49, 50, 55, 57, 58, 60}	-	22.21
18	-	-	{4, 7, 11, 13, 17, 18, 24, 29, 30, 31, 32, 34, 35, 36, 47, 49, 50, 51}	12.98
19	-	{10, 11, 12, 15, 17, 18, 21, 25, 29, 30, 34, 35, 45, 47, 49, 50, 52, 59, 60}	-	19.23
19	-	-	{4, 7, 11, 13, 17, 24, 29, 30, 31, 32, 34, 35, 36, 47, 49, 50, 51, 55, 58}	12.50
20	-	-	{4, 7, 11, 13, 17, 24, 29, 30, 31, 32, 34, 35, 36, 46, 47, 49, 50, 51, 55, 58}	12.50
21	-	-	{8, 11, 17, 18, 20, 21, 24, 27, 30, 31, 32, 35, 36, 39, 40, 43, 49, 50, 53, 60}	11.54
22	-	-	{8, 11, 17, 18, 19, 20, 21, 24, 27, 30, 31, 32, 35, 36, 39, 40, 43, 48, 49, 50, 53, 60}	11.06
26	-	-	{4, 8, 11, 17, 18, 19, 20, 21, 24, 27, 29, 30, 31, 32, 34, 35, 36, 39, 40, 43, 45, 47, 49, 50, 53, 55}	10.10
27	-	-	{4, 8, 11, 17, 18, 19, 20, 21, 24, 27, 29, 30, 31, 32, 34, 35, 36, 39, 40, 43, 45, 47, 49, 50, 53, 55, 60}	9.615

Archived multi-objective simulated annealing was adapted to cope with the feature selection problem. The modified AMOSA is compared with two multi-objective optimization algorithms: NSGA-II and DMS. In order to evaluate the feature subsets, fuzzy models are used.

Both NSGA-II and DMS outperformed the proposed modified AMOSA. NSGA-II is a population based multi-objective algorithm, which showed to be more efficient in approximating the Pareto front. DMS also progresses with the evolution of a set of nondominated solutions, and the greedy properties of the search mechanism granted a better performance for approximating the Pareto front. One of the key mechanisms in modified AMOSA is the creation of a new solution by using the neighborhood of the current solution, which is called perturbation. The results showed that this search mechanism is not effective. Thus, the perturbation scheme should be improved.

Acknowledgements The research in this work has been supported by the COST Action IC0702 STSMs. The research by Özlem Türkşen was partially supported by the TUBITAK (The Scientific and Technological Research Council of Turkey-code 2214-Research Project) which is gratefully acknowledged. This work was also supported by ISEL, byFundação para a Ciência e a Tecnologia (FCT), through IDMEC-IST under LAETA, and by a FCT grant SFRH/BPD/65215/2009, Ministério do Ensino Superior, da Ciência e da Tecnologia, Portugal.

References

1. Direct multisearch (dms) for multi-objective optimization. `http://www.mat.uc.pt/dms/` (2012)
2. Asuncion A, Newman D (2007) UCI machine learning repository. `http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html`
3. Bandyopadhyay S, Saha S, Maulik U, Deb K (2008) A simulated annealing-based multi-objective optimization algorithm: AMOSA. *IEEE Transactions on Evolutionary Computation* 12(3):269–283
4. van den Berg J, Kaymak U, van den Bergh WM (2002) Fuzzy classification using probability-based rule weighting. *Proc. IEEE Int. Conf. on Fuzzy Systems*, 2:991–996. IEEE Press, Piscataway
5. Bhatia S, Prakash P, Pillai G (2008) SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. *Proc. World Congress on Engineering and Computer Science*. San Francisco
6. Coello CC, Voldhuizen D, Lament G (eds.) (2002) *Evolutionary Algorithms for Solving Multi Objective Problems*. Kluwer Academic, New York
7. Conn AR, Scheinberg K, Vicente LN (2009) Introduction to derivative-free optimization. *MPS-SIAM Series on Optimization*. SIAM, Philadelphia
8. Cordon O, Herrera F, Jesus M, Magdalena L, Sanchez A, Villar P (2002) A multiobjective genetic algorithm for feature selection and data base learning in fuzzy-rule based classification systems. *Proc. 9th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002)*, 823–830. Annecy, France
9. Custódio AL, Madeira JFA, Vaz AIF, Vicente LN (2011) Direct multisearch for multiobjective optimization. *SIAM Journal on Optimization* 10–18
10. Deb K, ed. (2004) Multi Objective Optimization using Evolutionary Algorithms. J. Wiley & Sons, Chichester
11. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans. an Evolutionary Computation* 6(2):10–18
12. Ekbal A, Saha S, Garbe C (2010) Feature selection using multi-objective optimization for named entity recognition. *Proc. IEEE Int. Conf. on Pattern Recognition*, 1937–1940
13. Emmanouilidis C, ed. (2002) Evolutionary Multi-Objective Feature Selection and ROC Analysis with Application to Industrial Machinery Fault Diagnosis. *Evolutionary Methods for Design, Optimization and Control*. J. Wiley & Sons, Chichester
14. Emmanouilidis C, Hunter A, MacIntyre J (2000) A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. *Proc. Congress on Evolutionary Computation (CEC'2000)*, 823–830. San Diego
15. Emmanouilidis C, Hunter A, MacIntyre J, Cox C (2001) A multi-objective genetic algorithm approach to feature selection in neural and fuzzy modeling. *Evolutionary Optimization* 3(1):1–26

16. Gustafson D, Kessel W (1985) Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics* 15(1):116–132
17. Hamdani T, Won J, Alimi A, Karray F (2007) Multi-objective feature selection with NSGA-II. *Proc. of ICANNGA 2007*, 240–247
18. Huang B, Buckley B, Kechadi T (2010) Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. *Expert Systems with Applications* 37(5):3638–3646
19. Lac H, Stacey D (2005) Feature subset selection via multi-objective genetic algorithm. *Proc. IEEE Int. Joint Conf. on Neural Networks*, 1349–1354. IEEE Press, Piscataway
20. Michalewicz Z (1999) Genetic Algorithms + Data Structures = Evolution Programs, 3rd edition. Springer, Berlin
21. Miettinen K (1999) Nonlinear Multi-objective Optimization. Kluwer, New York
22. Nieto J, Alba E, Jourdan L, Talbi E (2009) Sensitivity and specificity based multi-objective approach for feature selection: Application to cancer diagnosis. *Information Processing Letters* 109:887–896
23. Oliveira L, Sabourin R, Bortolozzi F, Suen C (2002) Feature selection using multi-objective genetic algorithms for handwritten digit recognition. *Proc. 16th Int. Conf. on Pattern Recognation (ICPR' 02)*, 568–571. IEEE Press, Piscataway
24. Saha S, Ekbal A, Uryupina O, Poesio M (2011) Single and multi-objective optimization for feature selection in anaphora resolution. *Proc. 5th Int. Joint Conf. on Natural Language Processing*, 93–101
25. Sousa JMC, Kaymak U (2002) *Fuzzy Decision Making in Modeling and Control*. World Scientific and Imperial College, Singapore and UK
26. Srinivas N, Deb K (1994) Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation* 2(3):221–248
27. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modelling and control. *IEEE Trans. on Systems, Man and Cybernetics* 15(1):116–132
28. Unler A, Murat A (2010) A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research* 206:528–539
29. Vieira SM, Sousa JMC, Runkler TA (2010) Two cooperative ant colonies for feature selection using fuzzy models. *Expert Systems with Applications* 37(4):2714–2723
30. Wang C, Huang Y (2009) Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. *Expert Systems with Applications* 36:5900–5908
31. Wang X, Bandyopadhyay S, Xuan Z, Zhao X, Zhang M, Zhang X (2007) Prediction of transcription start sites based on feature selection using AMOSA. *Computational Systems Bioinformatics Conference*{ 6:183–193
32. Waqas K, Baig R, Ali S (2009) Feature subset selection using multi-objective genetic algorithms. *Proc. 13th IEEE Multitopic Conference*, 1–6.

Author Index